Schaum

PROBABILIDAD Y ESTADÍSTICA

SEGUNDA EDICIÓN

MURRAY R. SPIEGEL
JOHN SCHILLER
R. ALU SRINIVASAN

- 760 problemas totalmente resueltos.
- Terminología y notación modernas.
- Metodología fácil de entender.
- Perfecto para repasar antes de los exámenes.

Mc Graw Hill

Más de 30 millones de ejemplares vendidos

Teoría y problemas de

PROBABILIDAD Y ESTADÍSTICA

Segunda edición

MURRAY R. SPIEGEL, Ph. D.
Former Professor and Chairman of Mathematics
Rensselaer Polytechnic Institute
Hartford Graduate Center

JOHN J. SCHILLER, Ph. D.
Associate Professor of Mathematics
Temple University

R. ALU SRINIVASAN, Ph. D.
Professor of Mathematics
Temple University

Traducción
LIBIA PATRICIA PARDO MILLER
Bióloga Universidad de los Andes
Máster en Biología de la Universidad Bochum, Alemania

Revisión técnica
JACINTO LONDOÑO ORTIZ
Estadístico Universidad Nacional de Colombia
Magíster en Economía
Docente universitario
Pontificia Universidad Javeriana

Bogotá • Buenos Aires • Caracas • Guatemala • Lisboa • Madrid • México
New York • Panamá • San Juan • Santiago de Chile • São Paulo
Auckland • Hamburgo • Londres • Milán • Montreal • Nueva Delhi • París
San Francisco • San Luis • Singapur • Sidney • Tokio • Toronto

MURRAY R. SPIEGEL, fallecido, recibió su grado en física M. S. Y su Ph. D. en matemáticas en la Universidad de Cornell. Tuvo posiciones destacadas en la Universidad de Harvard, la Universidad de Columbia, el Instituto Politécnico de Oak Ridge y Rensselaer, y fue consultor matemático en varias compañías importantes. Su último puesto fue como profesor y director de matemáticas en el Instituto Politécnico, Hartford Graduate Center. Él estuvo interesado en la mayoría de las áreas de las matemáticas, especialmente en aquellas relacionadas con su aplicación a la solución de problemas en física e ingeniería. Fue autor de numerosas publicaciones y de catorce libros sobre varios temas matemáticos.

JOHN J. SCHILLER es profesor asociado de matemáticas en la Universidad de Temple. Recibió su Ph. D. en la Universidad de Pensilvania. Ha publicado artículos sobre investigaciones en las siguientes áreas: superficies de Riemann, matemáticas discretas y matemáticas para biología. Ha sido, también, coautor en varios textos de matemáticas.

R. ALU SRINIVASAN es profesor de matemáticas en la Universidad de Temple. Recibió su Ph. D. en la Universidad Wayne State, y ha publicado muchos artículos sobre probabilidad y estadística.

PROBABILIDAD Y ESTADÍSTICA
Segunda edición

McGraw-Hill Interamericana

No está permitida la reproducción total o parcial de este libro, ni su tratamiento informático, ni la transmisión de ninguna forma o por cualquier medio, ya sea electrónico, mecánico, por fotocopia, por registro u otros métodos, sin el permiso previo y por escrito de los titulares del Copyright.

DERECHOS RESERVADOS © 2003 respecto a la segunda edición por
McGRAW-HILL/INTERAMERICANA EDITORES, S.A. DE C.V.
A Subsidiary of The McGraw-Hill Companies, Inc.
Cedro Núm. 512, Col. Atlampa
Delegación Cuauhtémoc,
06450 México, D.F.
Miembro de la Cámara Nacional de la Industria Editorial Mexicana, Reg. Núm. 736

ISBN : 970-10-4231-X
ISBN : 958-41-0133-1 edición Colombiana
ISBN : 642-2923-2 edición anterior

Traducido de la segunda edición en inglés de: SCHAUM'S OUTLINE OF THEORY AND PROBLEMS OF PROBABILITY AND STATISTICS. Copyright © MM, by The McGraw-Hill Companies, Inc.
ISBN: 0-07-135004-7

Editora: Emma Ariza Herrera
Jefe de Producción: Consuelo E. Ruiz M.
Diagramación: Álvaro Garrido Madrid

Esta obra se termino de
imprimir en Febrero del 2006
en Orsa y Asociados S.A. de C.V.

4567890123 09876542103

Impreso en México Printed in Mexico

The McGraw-Hill Companies

PREFACIO A LA SEGUNDA EDICIÓN

La primera edición de *Probabilidad y estadística*, de la serie de Schaum, escrita por Murray R. Spiegel, apareció en 1975 y se han hecho 21 reimpresiones desde entonces. Gian-Carlo Rota describió en su libro *Indiscrete Thoughts* a su primo cercano, *Estadística* de Schaum, por el mismo autor, como la introducción impresa más clara a la estadística. De manera que fue con cierta reverencia y precaución que nos dimos a la tarea de revisar este libro. El principio que nos guió fue hacer los cambios necesarios para darle énfasis al texto en los temas contemporáneos. El tratamiento extensivo de conjuntos y material estándar introductorio en los textos de la década de los años 60 y de comienzo de los 70 se redujo considerablemente. La definición de una variable continua aleatoria es hoy en día estándar, y se hizo más énfasis en los valores *P* de las pruebas de hipótesis, desde que la tecnología hizo más sencillo determinar esos valores, lo cual provee de información más específica que si las pruebas cumplen con un nivel preestablecido de significancia. La tecnología también ha hecho posible eliminar las tablas de logaritmos. Se incluyó un capítulo de estadística no paramétrica para extender la aplicabilidad del texto sin aumentar su nivel. Se eliminaron algunos conjuntos de problemas, sólo en momentos en que se requerían pruebas para teoremas para los cuales no había elementos de ayuda de ninguna clase. En general creemos que el propósito principal de la primera edición –presentar una introducción moderna de la probabilidad y la estadística usando bases de cálculo– así como las características que hicieron de la primera edición tal éxito, se mantuvieron. Esperamos que esta edición sirva a un grupo de estudiantes aún mayor.

J. SCHILLER
R. A. SRINIVASAN

PREFACIO A LA PRIMERA EDICIÓN

La importancia de un tema tan importante y fascinante como la probabilidad empezó en la siglo XVII a través de los esfuerzos de matemáticos tales como Fermat y Pascal por resolver inquietudes concernientes a los juegos de azar. No fue sino hasta el siglo XX que se desarrolló una rigurosa teoría de las matemáticas basada en axiomas, definiciones y teoremas. Con el paso del tiempo, la probabilidad y la estadística se hicieron camino en campos que van desde la ciencia actuarial hasta la agricultura, y de los negocios hasta la medicina y la psicología. En muchos casos, las aplicaciones en sí mismas contribuyen al posterior desarrollo de la teoría.

El objeto de la estadística se originó mucho antes que el de la probabilidad y tuvo que ver principalmente con la recolección, organización y presentación de los datos en tablas y gráficos. Con el advenimiento de la probabilidad, se comprendió que la estadística podría usarse para establecer conclusiones válidas y tomar decisiones razonables con base en el análisis de datos, como en la teoría del muestreo y en predicciones.

El propósito de este libro es presentar una introducción moderna a la probabilidad y estadística usando como base el cálculo. Por conveniencia, el libro esta dividido en dos partes. La primera trata de probabilidad (y por sí misma puede usarse como una introducción a la materia), mientras que la segunda trata de estadística.

El libro está diseñado de manera que pueda usarse bien sea como libro de texto para un curso normal de probabilidad y estadística o como un complemento a cualquier texto corriente. Tiene también valor como libro de referencia para investigadores o para aquellos que estudian por su cuenta. El libro puede utilizarse para cursos de un año de duración, seleccionando los temas de interés o como un curso de un semestre.

Estoy muy agradecido con el albacea literario del fallecido Sir Ronald A. Fisher, F. R. S., con el Dr. Frank Yates, F. R. S. y con el grupo Longman Ltd., Londres, por el permiso para usar la tabla III de su libro *Statistical tables for Biological, Agricultural and Medical Research* (sexta edición, 1974). Quiero también aprovechar la oportunidad para agradecer a David Beckwith por su excelente trabajo de edición y a Nicola Monti por su trabajo artístico.

M. R. SPIEGEL

CONTENIDO

Parte I **PROBABILIDAD** **1**

CAPÍTULO 1 **Probabilidad básica** **3**

Experimentos aleatorios. Espacio de la muestra. Eventos. El concepto de probabilidad. Los axiomas de la probabilidad. Algunos teoremas importantes en probabilidad. Asignación de probabilidades. Probabilidad condicional. Teoremas de probabilidad condicional. Eventos independientes. Teorema o regla de Bayes. Análisis combinatorio. Principio fundamental de conteo. Diagramas de árbol. Permutaciones. Combinaciones. Coeficientes binomiales. Aproximación de Stirling a $n!$

CAPÍTULO 2 **Variables aleatorias y distribución de probabilidad** **38**

Variables aleatorias. Distribuciones de probabilidad discreta. Funciones de distribución para variables aleatorias. Funciones de distribución para variables aleatorias discretas. Variables aleatorias continuas. Interpretaciones gráficas. Distribuciones conjuntas. Variables aleatorias independientes. Cambio de variables. Distribuciones de probabilidad de funciones de variables aleatorias. Convoluciones. Distribuciones condicionales. Aplicaciones de la probabilidad geométrica.

CAPÍTULO 3 **Esperanza matemática** **82**

Definición de esperanza matemática. Funciones de variables aleatorias. Algunos teoremas sobre la esperanza. La varianza y la desviación estándar. Algunos teoremas sobre la varianza. Variables aleatorias estandarizadas. Momentos. Funciones generadoras de momento. Algunos teoremas sobre funciones generadoras de momento. Funciones características. Varianza para distribuciones conjuntas. Covarianza. Coeficiente de correlación. Esperanza condicional, varianza y momentos. Desigualdad de Chebyshev. Ley de los grandes números. Otras medidas de la tendencia central. Percentiles. Otras medidas de dispersión. Sesgo y curtosis.

CAPÍTULO 4 **Distribuciones de probabilidad especial** 117
La distribución binomial. Algunas propiedades de la distribución binomial. Ley de los grandes números para las pruebas de Bernoulli. La distribución normal. Algunas propiedades de la distribución normal. Relación entre las distribuciones binomial y normal. La distribución de Poisson. Algunas propiedades de la distribución de Poisson. Relación entre las distribuciones Poisson y binomial. Relación entre las distribuciones normal y Poisson. El teorema del límite central. La distribución multinomial. La distribución hipergeométrica. La distribución uniforme. La distribución de Cauchy. La distribución gamma. La distribución beta. La distribución chi cuadrado. La distribución *t* de Student. La distribución *F*. Relación entre las distribuciones chi cuadrado, *t* y *F*. La distribución normal bivariada. Diversas distribuciones.

Parte II **ESTADÍSTICA** 163

CAPÍTULO 5 **Teoría de muestreo** 165
Población y muestra. Inferencia estadística. Muestreo con y sin remplazo. Muestras aleatorias. Números aleatorios. Parámetros de la población. Estadístico muestral. Distribución de la muestra. La media de la muestra. Distribución muestral de medias. Distribución muestral de proporciones. Distribución muestral de diferencias y sumas. La varianza de la muestra. Distribución muestral de varianzas. Caso en que se desconoce la varianza de la población. Distribución muestral de las relaciones de varianzas. Otros estadísticos. Distribuciones de frecuencia. Distribuciones de frecuencia relativa. Cálculo de media, varianza y momentos para datos agrupados.

CAPÍTULO 6 **Teoría de la estimación** 209
Estimación insesgada y estimación eficiente. Estimación puntual y estimación por intervalo. Confiabilidad. Estimación por intervalo de confianza para parámetros de la población. Intervalo de confianza para medias. Intervalo de confianza para proporciones. Intervalo de confianza para diferencias y sumas. Intervalo de confianza para la varianza de una distribución normal. Intervalo de confianza para la relación entre varianzas. Estimación de máxima verosimilitud.

CAPÍTULO 7 **Pruebas de hipótesis y significancia** 228
Decisiones estadísticas. Hipótesis estadísticas. Hipótesis nula. Pruebas de hipótesis y significancia. Errores de tipo I y tipo II. Nivel de significancia. Pruebas que involucran la distribución normal.

Pruebas de una y dos colas. El valor *P*. Pruebas especiales de significancia para muestras grandes. Pruebas especiales de significancia para muestras pequeñas. Relación entre la teoría de la estimación y la prueba de hipótesis. Curvas de operación características. Potencia de una prueba. Gráficos de control de calidad. Ajuste de las distribuciones teóricas a distribuciones de frecuencia muestrales. La prueba de chi cuadrado para la bondad de ajuste. Tablas de contingencia. Corrección de Yates por la continuidad. Coeficiente de contingencia.

CAPÍTULO 8 **Curva de ajuste, regresión y correlación** **285**

Curva de ajuste. Regresión. Método de los mínimos cuadrados. Recta de mínimos cuadrados. Recta de mínimos cuadrados en términos de varianza muestral y covarianza. Parábola de mínimos cuadrados. Regresión múltiple. Error estándar del estimado. Coeficiente de correlación lineal. Coeficiente de correlación generalizado. Correlación del rango. Interpretación probabilística de la regresión. Interpretación probabilística de la correlación. Teoría muestral de la regresión. Teoría muestral de la correlación. Correlación y dependencia.

CAPÍTULO 9 **Análisis de varianza** **335**

El propósito del análisis de varianza. Clasificación a una vía o experimentos de un factor. Variación total. Variación dentro de tratamientos. Variación entre tratamientos. Métodos abreviados para obtener variaciones. Modelo matemático lineal para el análisis de varianza. Valores esperados de las variaciones. Distribuciones de las variaciones. La prueba *F* para la hipótesis nula de medias iguales. Tabla de análisis de varianza. Modificaciones para números desiguales de observaciones. Clasificación a dos vías o experimentos de dos factores. Notación para experimentos de dos factores. Variaciones para experimentos de dos factores. Análisis de varianza para experimentos de dos factores. Experimentos de dos factores con replicación. Diseño experimental.

CAPÍTULO 10 **Pruebas no paramétricas** **372**

Introducción. La prueba del signo. La prueba *U* de Mann-Whitney. La prueba *H* de Kruskal-Wallis. La prueba *H* corregida para empates. La prueba de corridas para aleatoriedad. Aplicaciones adicionales de la prueba de corridas. Correlación de rango de Spearman.

Apéndice A **397**

Temas matemáticos

Apéndice B 400
Ordenadas (y) de la curva normal estándar en z

Apéndice C 401
Áreas bajo la curva normal estándar de 0 a z

Apéndice D 402
Valores percentiles t_p para la distribución t de Student con ν grados de libertad

Apéndice E 403
Valores percentiles de χ^2_p para la distribución chi cuadrado con ν grados de libertad

Apéndice F 404
Percentiles 95 (niveles de 0.05), $F_{0.95}$, para la distribución F
Percentiles 99 (niveles de 0.01), $F_{0.99}$, para la distribución F

Apéndice G 406
Valores de $e^{-\lambda}$

Apéndice H 407
Números aleatorios

ÍNDICE DE TEMAS 409

ÍNDICE DE PROBLEMAS RESUELTOS 415

PARTE I

Probabilidad

CAPÍTULO 1

Probabilidad básica

EXPERIMENTOS ALEATORIOS

Todos estamos familiarizados con la importancia de los experimentos en ciencias e ingeniería. La experimentación es útil porque si suponemos que llevamos a cabo ciertos experimentos bajo condiciones esencialmente idénticas, llegaremos a los mismos resultados. En estas circunstancias, estamos en capacidad de controlar el valor de las variables que afectan el resultado del experimento.

Sin embargo, en algunos experimentos, no somos capaces de indagar o controlar el valor de determinadas variables, de manera que el resultado cambiará de un experimento al otro, a pesar de que la mayoría de las condiciones son las mismas. Estos experimentos se describen como *aleatorios*. Los siguientes son algunos ejemplos.

EJEMPLO 1.1. Si lanzamos una moneda, el resultado del experimento será "sello", simbolizado por S (o 0), o "cara", simbolizado por C (o 1), es decir, uno de los elementos del conjunto $\{S, C\}$ o $\{0, 1\}$.

EJEMPLO 1.2. Si lanzamos un dado, el resultado del experimento será uno de los números del conjunto {1, 2, 3, 4, 5, 6}.

EJEMPLO 1.3. Si lanzamos dos veces una moneda, hay cuatro resultados posibles, tal como se indica por $\{CC, CS, SC, SS\}$, es decir, dos caras, cara en primer lugar y sello en segundo, etc.

EJEMPLO 1.4. Si estamos fabricando tornillos con una máquina, el resultado del experimento es que alguno puede ser defectuoso. Así, cuando se fabrica un tornillo, será un miembro del conjunto {defectuoso, no defectuoso}.

EJEMPLO 1.5. Si un experimento consiste en medir la vida útil de bombillas eléctricas producidas por una compañía, entonces el resultado del experimento es el tiempo t en horas que se encuentra en algún intervalo, por ejemplo, $-0 \leqq t \leqq 4000$ — donde suponemos que ninguna bombilla dura más de 4000 horas.

ESPACIOS DE LA MUESTRA

Un conjunto S que consta de todos los resultados posibles de un experimento aleatorio se llama *espacio muestral*, y cada resultado se denomina *punto muestral*. Con frecuencia habrá más de un espacio muestral que puede describir los resultados de un experimento, pero generalmente habrá uno que provee la mayor información.

EJEMPLO 1.6. Si lanzamos un dado, un espacio muestral, o conjunto de todos los resultados posibles, está dado por {1, 2, 3, 4, 5, 6} mientras que otro es {impar, par}. Sin embargo, es claro que el último no sería adecuado para determinar, por ejemplo, si un resultado es divisible por 3.

Con frecuencia es útil representar gráficamente el espacio muestral. En tales casos es aconsejable usar números en lugar de letras, cuando sea posible.

EJEMPLO 1.7. Si lanzamos dos veces una moneda, y usamos 0 para representar sellos y 1 para representar caras, el espacio muestral (*véase* el ejemplo 1.3) puede dibujarse por puntos como en la Figura 1-1 donde, por ejemplo, (0, 1) representa sellos en la primera lanzada y caras en la segunda, es decir, *SC*.

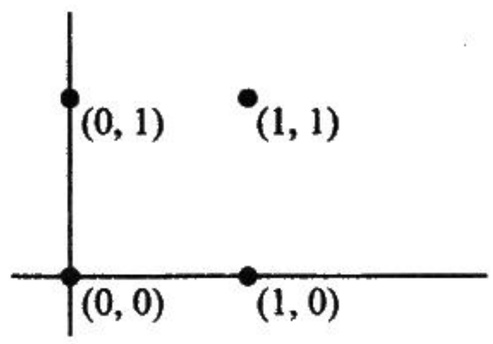


Fig. 1-1

Si un espacio muestral tiene un número finito de puntos, como en el ejemplo 1.7, se llama *espacio muestral finito*. Si tiene tantos puntos como los números naturales 1, 2, 3, . . ., se denomina *espacio muestral infinito contable*. Si tiene tantos puntos como hay en algunos intervalos del eje x, tal como en $0 \leqq x \leqq 1$, se llama *espacio muestral infinito no contable*. Un espacio muestral que es finito o infinito contable con frecuencia se denomina *espacio muestral discreto*, mientras que uno infinito no contable se llama *espacio muestral no discreto*.

EVENTOS

Un *evento* es un subconjunto A del espacio muestral S, es decir, un conjunto de resultados posibles. Si el resultado de un experimento es un elemento de A, decimos que *el evento A ocurrió*. Un evento que consta de un punto sencillo de S se denomina con frecuencia un *evento simple* o *elemental*.

EJEMPLO 1.8. Si lanzamos una moneda dos veces, el evento de que sólo resulte una cara es el subconjunto del espacio muestral que consta de los puntos (0, 1) y (1, 0), como se indica en la figura 1-2.

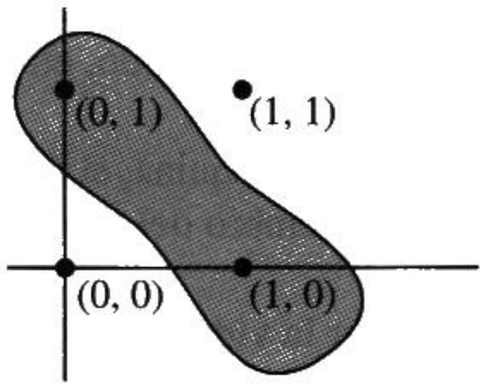


Fig. 1-2

Como eventos particulares tenemos al mismo S, el cual es el *evento cierto* o *seguro*, dado que un elemento S debe ocurrir, y el conjunto vacío $\varnothing$, que se denomina el *evento imposible* porque un elemento de $\varnothing$ no puede ocurrir.

Usando operaciones de conjuntos sobre eventos en S, podemos obtener otros eventos en S. Por ejemplo, si A y B son eventos, entonces

1. $A \cup B$ es el evento "A o B o ambos". $A \cup B$ se denomina la *unión* de A y B.
2. $A \cap B$ es el evento "A y B". $A \cap B$ se denomina la *intersección* de A y B.
3. A' es el evento "no A". A' se denomina el complemento de A.
4. $A - B = A \cap B'$ es el evento "A pero no B". En particular, $A' = S - A$.

Si los conjuntos que corresponden a los eventos A y B son disjuntos, es decir, $A \cap B = \varnothing$, decimos que los eventos son *mutuamente excluyentes*. Esto significa que no pueden ocurrir ambos. Decimos que una colección de eventos $A_1, A_2, \ldots A_n$ es mutuamente excluyente si cada par en la colección es mutuamente excluyente.

EJEMPLO 1.9. En relación con el experimento de lanzar dos veces una moneda, digamos que A es el evento "al menos ocurre una cara" y B el evento "el segundo lanzamiento es un sello". Entonces, $A = \{CS, SC, CC\}$, $B = \{CS, SS\}$, de manera que tenemos:

$$A \cup B = \{CS, SC, CC, SS\} = S \qquad A \cap B = \{CS\}$$
$$A' = \{SS\} \qquad A - B = \{SC, CC\}$$

EL CONCEPTO DE PROBABILIDAD

En cualquier experimento aleatorio hay siempre incertidumbre sobre si ocurrirá un evento en particular. Como una medida de la *oportunidad*, o *probabilidad*, con que esperamos que ocurra cierto evento, es conveniente asignar un número entre 0 y 1. Si estamos seguros de que tal evento ocurrirá, decimos que tiene 100% de probabilidad o 1, pero si estamos seguros de que el evento no ocurrirá, decimos que su probabilidad es cero. Si, por ejemplo, la probabilidad es $\frac{1}{4}$, debemos decir que la oportunidad de que ocurra es 25%, y 75% de que no ocurra. Igualmente, podemos decir que la probabilidad de no ocurrencia es 75% contra 25%, o 3 a 1.

Hay dos procedimientos importantes mediante los cuales podemos calcular la probabilidad de un evento.

1. **EL ENFOQUE CLÁSICO.** Si un evento puede ocurrir en h maneras diferentes de un número total de n maneras posibles, todos ellos son igualmente posibles. Entonces, la probabilidad del evento es h/n.

EJEMPLO 1.10. Supongamos que queremos saber la probabilidad de que ocurra cara en el lanzamiento sencillo de una moneda. Dado que hay dos maneras igualmente probables como puede caer una moneda, a saber: cara o sello (suponiendo que no se pierda o caiga sobre el borde), y que de esas dos maneras puede aparecer cara de una sola forma, deducimos que la probabilidad es $1/2$. Para llegar a esto debemos suponer que la moneda es *balanceada*, es decir, que no está *cargada* de alguna manera.

2. **EL ENFOQUE FRECUENTISTA.** Si después de n repeticiones de un experimento, donde n es muy grande, se observa que un evento ocurre h veces, entonces la probabilidad de dicho evento es h/n. Esto también se denomina la *probabilidad empírica* de un evento.

EJEMPLO 1.11. Si lanzamos 1000 veces una moneda y encontramos que 532 veces obtenemos caras, estimamos que la probabilidad de que ocurra cara es de $532/1000 = 0{,}532$.

Tanto el enfoque clásico como el frecuentista presentan serios inconvenientes. El primero porque las palabras "igualmente probable" son vagas y el segundo porque el "número grande" es vago. Debido a estas dificultades, los matemáticos se han regido por el *enfoque axiomático* de la probabilidad.

LOS AXIOMAS DE LA PROBABILIDAD

Supongamos que tenemos un espacio muestral S. Si S es discreto, todos los subconjuntos corresponden a eventos e inversamente, pero si S no es discreto, sólo los subconjuntos especiales (llamados *medibles*) corresponden a eventos. Para cada evento A en la clase C de eventos, asociamos un número real $P(A)$. Entonces P se denomina la *función de probabilidad*, y $P(A)$ la probabilidad del evento A, si se cumplen los siguientes axiomas.

Axioma 1. Para cada evento A en la clase C,

$$P(A) \geqq 0 \tag{1}$$

Axioma 2. Para el evento cierto o seguro S en la clase C,

$$P(S) = 1 \tag{2}$$

Axioma 3. Para cualquier número de eventos mutuamente excluyentes $A_1, A_2, \ldots$, en la clase C,

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots \tag{3}$$

En particular, para dos eventos mutuamente excluyentes A_1, A_2,

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) \tag{4}$$

ALGUNOS TEOREMAS IMPORTANTES SOBRE PROBABILIDAD

A partir de los axiomas anteriores podemos probar varios teoremas sobre probabilidad que son importante para el estudio posterior.

Teorema 1-1: Si $A_1 \subset A_2$, entonces $P(A_1) \leqq P(A_2)$ y $P(A_2 - A_1) = P(A_2) - P(A_1)$.

Teorema 1-2: Para todo evento A,

$$0 \leqq P(A) \leqq 1, \tag{5}$$

o sea, la probabilidad está entre 0 y 1.

Teorema 1-3:

$$P(\varnothing) = 0 \tag{6}$$

es decir, el evento imposible tiene probabilidad cero.

Teorema 1-4: Si A' es el complemento de A, entonces

$$P(A') = 1 - P(A) \tag{7}$$

Teorema 1-5: Si $A = A_1 \cup A_2 \cdots \cup A_n$, donde $A_1, A_2, \ldots, A_n$ son eventos mutuamente excluyentes, entonces

$$P(A) = P(A_1) + P(A_2) + \cdots + P(A_n) \tag{8}$$

En particular, si $A = S$, el espacio muestral, entonces

$$P(A_1) + P(A_2) + \cdots + P(A_n) = 1 \tag{9}$$

Teorema 1-6: Si A y B son dos eventos cualesquiera, entonces

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{10}$$

De manera más general, si A_1, A_2, A_3 son tres eventos cualesquiera, entonces

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_2 \cap A_3) - P(A_3 \cap A_1) + P(A_1 \cap A_2 \cap A_3) \tag{11}$$

También se pueden hacer generalizaciones para n eventos.

Teorema 1-7: Para cualesquiera eventos A y B,

$$P(A) = P(A \cap B) + P(A \cap B') \tag{12}$$

Teorema 1-8: Si un evento A debe dar como resultado la ocurrencia de uno de los eventos mutuamente excluyentes $A_1, A_2, \ldots, A_n$, entonces

$$P(A) = P(A \cap A_1) + P(A \cap A_2) + \cdots + P(A \cap A_n) \tag{13}$$

ASIGNACIÓN DE PROBABILIDADES

Si un espacio muestral S consta de un número finito de resultados $a_1, a_2, \ldots, a_n$, entonces según el teorema 1-5,

$$P(A_1) + P(A_2) + \cdots + P(A_n) = 1 \tag{14}$$

donde $A_1, A_2, \ldots, A_n$ son eventos elementales dados por $A_i = \{a_i\}$.

Entonces, podemos escoger arbitrariamente cualquier número no negativo para las probabilidades de esos eventos sencillos siempre y cuando (*14*) se satisfaga. En particular, si suponemos que hay *probabilidades iguales* para todos los eventos sencillos, entonces

$$P(A_k) = \frac{1}{n}, \qquad k = 1, 2, \ldots, n \tag{15}$$

y si A es un evento cualquiera hecho de h eventos sencillos, tenemos

$$P(A) = \frac{h}{n} \tag{16}$$

Esto equivale al enfoque clásico de la probabilidad dado en la página 5. Podemos, desde luego, usar otros procedimientos para asignar probabilidades, tales como el enfoque frecuentista de la página 5.

La asignación de probabilidades provee un *modelo matemático*, cuyo éxito debe comprobarse de manera experimental, tal como se hace con las teorías de la física y otras ciencias.

EJEMPLO 1.12. Un dado se lanza una vez. Encuentre la probabilidad de un 2 o un 5.

El espacio muestral es $S = \{1, 2, 3, 4, 5, 6\}$. Si asignamos probabilidades iguales a los puntos muestrales, es decir, si suponemos que el dado es balanceado, entonces

$$P(1) = P(2) = \cdots = P(6) = \frac{1}{6}$$

El evento de que salga tanto un 2 como un 5 se indica por $2 \cup 5$. Por consiguiente,

$$P(2 \cup 5) = P(2) + P(5) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

PROBABILIDAD CONDICIONAL

Sean A y B dos eventos (figura 1-3) tales que $P(A) > 0$. Denótese por $P(B \mid A)$ la probabilidad de B dado que ocurrió A. Puesto que sabemos que ocurrió A, éste se convierte en el nuevo espacio muestral remplazando al original S. A partir de esto llegamos a la definición

$$P(B \mid A) \equiv \frac{P(A \cap B)}{P(A)} \tag{17}$$

o

$$P(A \cap B) \equiv P(A)P(B \mid A) \tag{18}$$

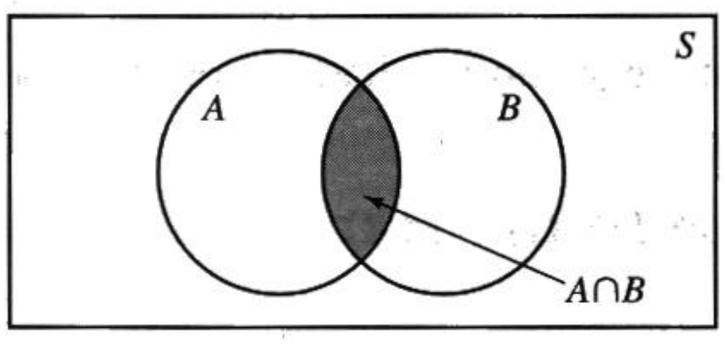


Fig. 1-3

En palabras, (*18*) dice que la probabilidad de que tanto A como B ocurran es igual a la probabilidad de que A ocurra por la probabilidad de que ocurra B dado que A ya ocurrió. Podemos llamar $P(B/A)$ como la *probabilidad condicional* de B dado A, es decir, la probabilidad de que B ocurra dado que A ocurrió. Es fácil mostrar que la probabilidad condicional satisface los axiomas de la página 6.

EJEMPLO 1.13. Encuentre la probabilidad de que en el lanzamiento sencillo de un dado resulte un número menor que 4 si (*a*) no se da otra información y (*b*) el dado cayó en número impar.

(*a*) Digamos que B indica el evento {menor que 4}. Dado que B es la unión de los eventos 1, 2 ó 3, vemos a partir del teorema 1-5 que

$$P(B) = P(1) + P(2) + P(3) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

suponiendo probabilidades iguales para los puntos muestrales.

(*b*) Al decir que A es el evento {números impares}, vemos que $P(A) = \frac{3}{6} = \frac{1}{2}$. También $P(A \cap B) = \frac{2}{6} = \frac{1}{3}$. Entonces

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{1/3}{1/2} = \frac{2}{3}$$

Por lo tanto, saber que el lanzamiento da como resultado un número impar eleva la probabilidad de 1/2 a 2/3.

TEOREMAS DE PROBABILIDAD CONDICIONAL

***Teorema 1-9*:** Para tres eventos cualesquiera A_1, A_2, A_3, tenemos

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \tag{19}$$

En palabras, la probabilidad de que A_1 y A_2 y A_3 ocurran es igual a la probabilidad de que A_1 ocurra por la probabilidad de que A_2 ocurra, dado que A_1 ocurrió por la probabilidad de que A_3 ocurra dado que tanto A_1 como A_2 ocurrieron. El resultado se generaliza fácilmente a n eventos.

***Teorema 1-10*:** Si un evento A debe originar uno de los eventos mutuamente excluyentes $A_1, A_2, \ldots, A_n$, entonces

$$P(A) = P(A_1)P(A \mid A_1) + P(A_2)P(A \mid A_2) + \cdots + P(A_n)P(A \mid A_n) \tag{20}$$

EVENTOS INDEPENDIENTES

Si $P(B \mid A) = P(B)$, es decir, que la probabilidad de que ocurra B no está afectada por la ocurrencia o no de A, entonces decimos que A y B son *eventos independientes*. Esto equivale a

$$P(A \cap B) = P(A)P(B) \tag{21}$$

como vimos en (*18*). Inversamente, si se cumple (*21*), entonces A y B son independientes.

Decimos que tres eventos A_1, A_2, A_3 son *independientes* si son independientes por parejas

$$P(A_j \cap A_k) = P(A_j)P(A_k) \qquad j \neq k \qquad \text{donde} \quad j,k = 1,2,3 \tag{22}$$

y

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3) \tag{23}$$

Note que ni (*22*) ni (*23*) son suficientes por sí mismos. La independencia de más de tres eventos se define fácilmente.

TEOREMA O REGLA DE BAYES

Suponga que $A_1, A_2, \ldots, A_n$ son eventos mutuamente excluyentes cuya unión es el espacio muestral S, es decir, que debe ocurrir uno de los eventos. Entonces si A es un evento cualquiera, tenemos el siguiente teorema importante:

***Teorema 1-11* (Regla de Bayes):**

$$P(A_k \mid A) = \frac{P(A_k)P(A \mid A_k)}{\sum_{j=1}^{n} P(A_j)P(A \mid A_j)} \tag{24}$$

Esto nos permite encontrar las probabilidades de diferentes eventos $A_1, A_2, \ldots, A_n$ que pueden *hacer* que ocurra A. Por esta razón, con frecuencia se hace referencia al teorema de Bayes como el *teorema de la probabilidad de las causas*.

ANÁLISIS COMBINATORIO

En muchos casos el número de puntos muestrales en un espacio muestral no es muy grande, de manera que no es difícil la enumeración directa o el conteo de los puntos muestrales necesario para obtener las probabilidades. Sin embargo, los problemas surgen donde el conteo directo se vuelve una imposibilidad práctica. En tales casos, se hace uso del *análisis combinatorio*, el cual se puede también denominar una *manera sofisticada de contar*.

PRINCIPIO FUNDAMENTAL DE CONTEO. DIAGRAMAS DE ÁRBOL

Si algo puede hacerse de n_1 maneras diferentes, y luego de ello se puede hacer una segunda cosa en n_2 maneras distintas, . . ., y finalmente se puede hacer una k-ésima en n_k formas diferentes, entonces todas las k cosas pueden llevarse a cabo en un orden específico en $n_1 * n_2 * \cdots * n_k$ formas diferentes.

EJEMPLO 1.14. Si un hombre tiene 2 camisas y 4 corbatas, entonces él tiene $2 \cdot 4 = 8$ formas de escoger una camisa y una corbata.

En conexión con el principio anterior se usa un diagrama llamado *diagrama de árbol* debido a su apariencia (figura 1-4).

EJEMPLO 1.15. En el diagrama de árbol de la figura 1-4 se indican las maneras diferentes como se puede escoger primero una camisa y luego una corbata. Las camisas se representan con S_1, S_2 y las corbatas con T_1, T_2, T_3 y T_4.

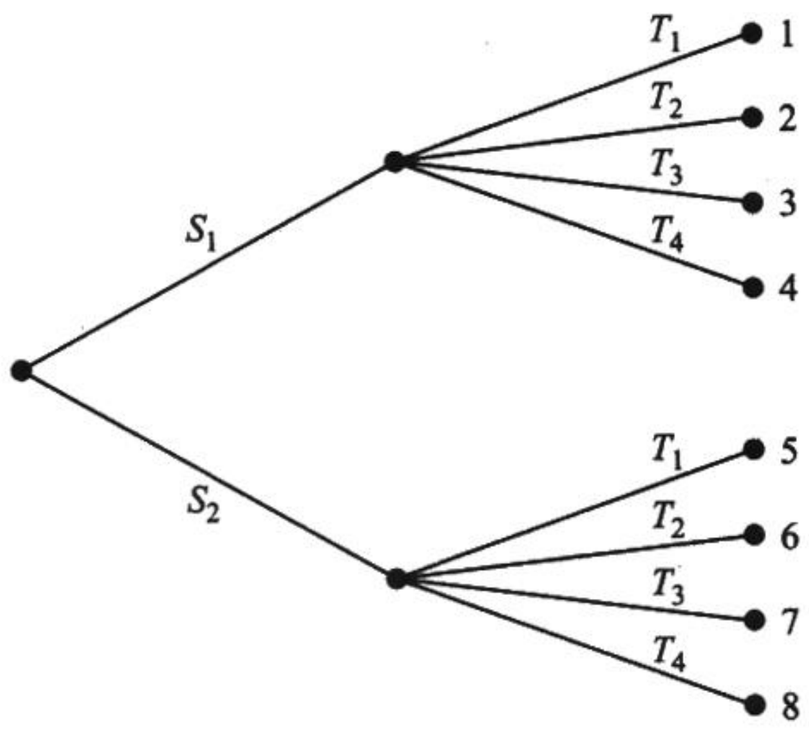


Fig. 1-4

PERMUTACIONES

Suponga que hay n objetos diferentes y queremos ordenar r de estos objetos en línea. Puesto que hay n maneras de escoger el primer objeto y luego $n - 1$ maneras de escoger el segundo, . . ., y finalmente $n - r + 1$ maneras de escoger el r-ésimo, a partir del principio fundamental del conteo se deduce que el número de arreglos diferentes o permutaciones, como se las nombra con frecuencia, está dado por

$$ {}_nP_r = n(n-1)(n-2)\cdots(n-r+1) \tag{25} $$

donde se observa que el producto tiene r factores. Llamamos ${}_nP_r$ al *número de permutaciones de n objetos tomados r a la vez.*

En el caso particular donde $r = n$, (25) se obtiene

$$ {}_nP_n = n(n-1)(n-2)\cdots 1 = n! \tag{26} $$

el cual se llama n factorial. Podemos escribir (25) en términos de factorial como

$$ {}_nP_r = \frac{n!}{(n-r)!} \tag{27} $$

Si $r = n$, vemos que (27) y (26) se satisfacen sólo si tenemos $0! = 1$, lo cual tomaremos como la definición de 0!

EJEMPLO 1.16. El número de arreglos diferentes, o permutaciones, que constan de 3 letras cada una, que pueden formarse a partir de 7 letras A, B, C, D, E, F, G es

$$ {}_7P_3 = \frac{7!}{4!} = 7 \cdot 6 \cdot 5 = 210 $$

Suponga que un conjunto consta de n objetos de los cuales n_1 son de un tipo (es decir, que no pueden distinguirse entre sí), n_2 son de otro tipo, . . ., n_k son de un k-ésimo tipo. Aquí, por supuesto, $n = n_1 + n_2 + \cdots + n_k$. Entonces el número de permutaciones diferentes del objeto es

$$ {}_nP_{n_1,n_2,\ldots,n_k} = \frac{n!}{n_1!n_2!\cdots n_k!} \tag{28} $$

Véase el problema 1.25.

EJEMPLO 1.17. El número de permutaciones diferentes de 11 letras de la palabra *M I S S I S S I P P I*, la cual consta de 1 *M*, 4 *I*, 4 *S* y 2 *P*, es

$$ \frac{11!}{1!4!4!2!} = 34{,}650 $$

COMBINACIONES

En una permutación nos interesa el orden de los objetos. Por ejemplo, *abc* es una permutación diferente de *bca*. Sin embargo, en mucho problemas sólo nos interesa la selección de los objetos sin tener en cuenta su orden. Este tipo de selecciones reciben el nombre de *combinaciones*. Por ejemplo, *abc* y *bca* son la misma combinación.

El número total de combinaciones de r objetos seleccionados de n (también conocido como *combinaciones* de n tomadas en r) se denota por ${}_nC_r$ o $\binom{n}{r}$. Tenemos (*véase* problema 1.27)

$$ \binom{n}{r} = {}_nC_r = \frac{n!}{r!(n-r)!} \tag{29} $$

Tambiém puede escribirse

$$ \binom{n}{r} = \frac{n(n-1)\cdots(n-r+1)}{r!} = \frac{{}_nP_r}{r!} \tag{30} $$

Es fácil mostrar que

$$ \binom{n}{r} = \binom{n}{n-r} \quad \text{o} \quad {}_nC_r = {}_nC_{n-r} \tag{31} $$

EJEMPLO 1.18. El número de maneras como pueden escogerse 3 de un total de 8 cartas diferentes es

$$ {}_8C_3 = \binom{8}{3} = \frac{8\cdot 7\cdot 6}{3!} = 56 $$

COEFICIENTES BINOMIALES

Los números (*29*) con frecuencia se llaman *coeficientes binomiales* porque provienen de la *expansión binomial*

$$ (x+y)^n = x^n + \binom{n}{1}x^{n-1}y + \binom{n}{2}x^{n-2}y^2 + \cdots + \binom{n}{n}y^n \tag{32} $$

Ellos tienen muchas propiedades interesantes.

EJEMPLO 1.19.

$$ \begin{aligned} (x+y)^4 &= x^4 + \binom{4}{1}x^3y + \binom{4}{2}x^2y^2 + \binom{4}{3}xy^3 + \binom{4}{4}y^4 \\ &= x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4 \end{aligned} $$

APROXIMACIÓN DE STIRLING A $n!$

Cuando n es grande, no es práctica la evaluación directa de $n!$ En tales casos se puede usar la siguiente fórmula de aproximación

$$n! \sim \sqrt{2\pi n}\, n^n e^{-n} \tag{33}$$

donde $e = 2.71828\ldots$, que es la base de los logaritmos naturales. El símbolo $\sim$ en (*33*) significa que la relación del lado izquierdo se aproxima a 1, a medida que $n \to \infty$.

La tecnología de los computadores ha eclipsado en buena medida el valor de la fórmula de Stirling para el cálculo numérico, pero la aproximación continúa teniendo importancia para estimaciones teóricas (*véase* el apéndice A).

Problemas resueltos

EXPERIMENTOS ALEATORIOS, ESPACIOS MUESTRALES Y EVENTOS

1.1. Se saca una carta al azar de una baraja de 52. Describa el espacio muestral si (*a*) no se considera el palo, (*b*) se considera el palo.

(*a*) Si no tenemos en consideración el palo, el espacio muestral consta de as, dos, . . ., diez, jota, reina, rey, y se puede indicar como $\{1, 2, \ldots, 13\}$.

(*b*) Si tenemos en consideración el palo, el espacio muestral consta de as de corazones, picas, diamantes y tréboles; . . . ; rey de corazones, picas, diamantes y tréboles. Si describimos corazones, picas, diamantes y tréboles, respectivamente por 1, 2, 3, 4, podemos, por ejemplo, indicar la jota de picas como (11, 2). El espacio muestral consta de 52 puntos, como se muestra en la figura 1-5.

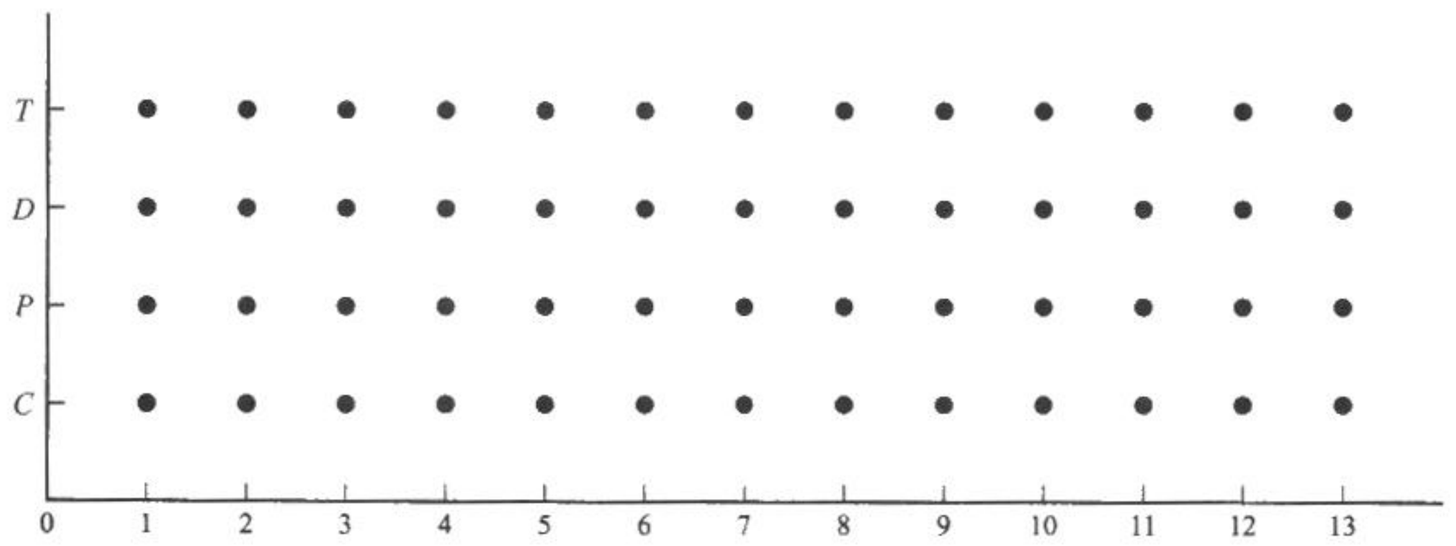


Fig. 1-5

1.2. En relación con el problema del experimento 1.1, digamos que A es el evento {sacar un rey} o simplemente {rey} y B es el evento {sacar un trébol} o simplemente {trébol}. Describa los eventos (*a*) $A \cup B$, (*b*) $A \cap B$, (*c*) $A \cup B'$, (*d*) $A' \cup B'$, (*e*) $A - B$, (*f*) $A' - B'$, (*g*) $(A \cap B) \cup (A \cap B')$.

(*a*) $A \cup B$ = {rey o trébol (o ambos, o sea, rey de tréboles)}.

(*b*) $A \cap B$ = {rey y trébol} = {rey de tréboles}.

(c) Puesto que B = {trébol}, B' = {no trébol} = {corazón, diamante, pica}.
Entonces, $A \cup B'$ = {rey de corazones, diamantes o picas}.

(*d*) $A' \cup B'$ = {no rey o no trébol} = {no rey de tréboles} = {cualquier carta diferente a rey de tréboles}.
Esto también se puede demostrar por $A' \cup B' = (A \cap B)'$ y usando (*b*).

(*e*) $A - B$ = {rey pero no trébol}.
Es lo mismo que $A \cap B'$ = {rey y no trébol}.

(*f*) $A' - B' = \{$no rey y no "no trébol"$\} = \{$no rey y trébol$\} = \{$cualquier trébol menos el rey$\}$.
Esto puede verse también teniendo en cuenta que $A' - B' = A' \cap (B')' = A' \cap B$.

(*g*) $(A \cap B) \cup (A \cap B') = \{$(rey y trébol) o (rey y no trébol)$\} = \{$rey$\}$.
Esto puede verse también teniendo en cuenta que $(A \cap B) \cup (A \cap B') = A$.

1.3. Use la figura 1-5 para describir los eventos (*a*) $A \cup B$, (*b*) $A' \cap B'$.

Los eventos requeridos se indican en la figura 1-6. De manera similar, todos los eventos del problema 1.2 se pueden indicar también por este tipo de diagramas. Se debe observar de la figura 1.6 que $A' \cap B'$ es complemento de $A \cup B$.

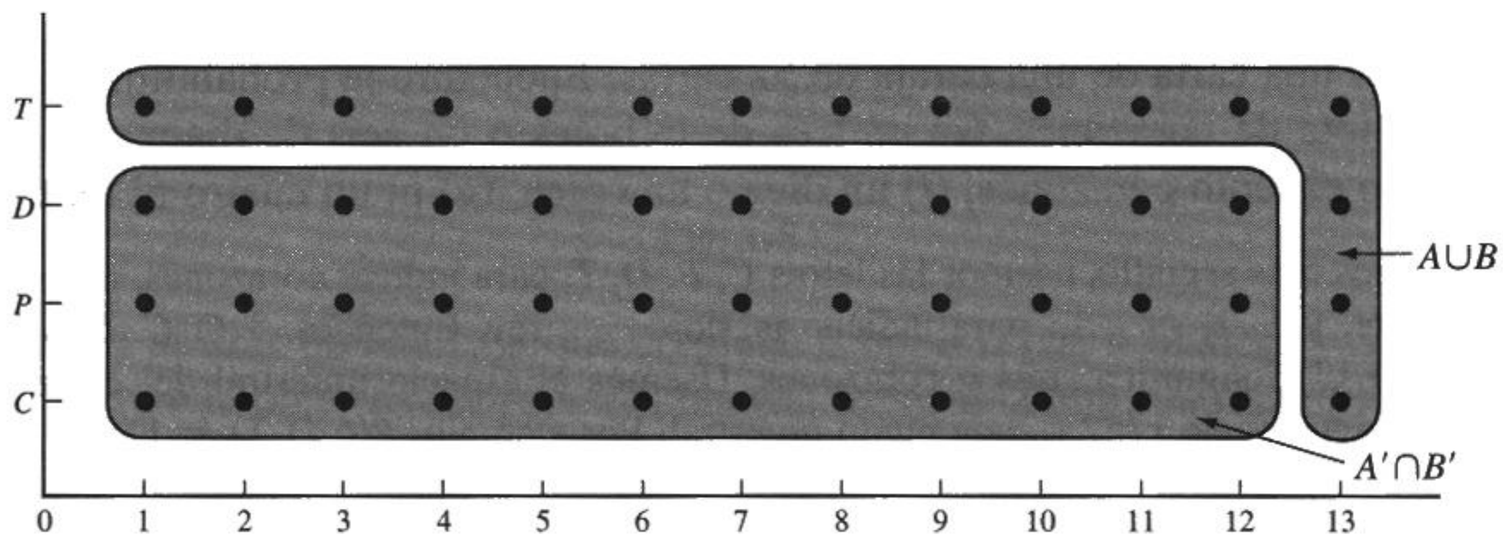


Fig. 1-6

TEOREMAS EN PROBABILIDAD

1.4. Demuestre (*a*) El teorema 1-1, (*b*) El teorema 1-2, (*c*) El teorema 1-3, de la página 6.

(*a*) Tenemos que $A_2 = A_1 \cup (A_2 - A_1)$ donde A_1 y $A_2 - A_1$ son mutuamente excluyentes. Entonces según el axioma 3, página 6:

$$P(A_2) = P(A_1) + P(A_2 - A_1)$$

de manera que
$$P(A_2 - A_1) = P(A_2) - P(A_1)$$

Dado que $P(A_2 - A_1) \geqq 0$ según el axioma 1, página 6, tenemos que $P(A_2) \geqq P(A_1)$.

(b) Según el axioma 1 ya sabemos que $P(A) \geqq 0$. Para probar que $P(A) \leqq 1$, debemos primero notar que $A \subset S$. Entonces, a partir del teorema 1-1 [parte (a)] y del axioma 2,

$$P(A) \leqq P(S) = 1$$

(c) Tenemos que $S = S \cup \varnothing$. Puesto que $S \cap \varnothing = \varnothing$, a partir del axioma 3 decimos que

$$P(S) = P(S) + P(\varnothing) \qquad \text{o} \qquad P(\varnothing) = 0$$

1.5. Demuestre (*a*) El teorema 1-4, (*b*) El teorema 1-6

(*a*) Tenemos que $A \cup A' = S$. Entonces, dado que $A \cap A' = \varnothing$, tenemos

$$P(A \cup A') = P(S) \qquad \text{o} \qquad P(A) + P(A') = 1$$

es decir
$$P(A') = 1 - P(A)$$

(*b*) A partir del diagrama de Venn de la figura 1-7 tenemos que,

(*1*)
$$A \cup B = A \cup [B(A \cap B)]$$

Entonces, dado que los conjuntos A y $B - (A \cap B)$ son mutuamente excluyentes, tenemos, usando el axioma 3 y el teorema 1-1,

$$P(A \cup B) = P(A) + P[B - (A \cap B)]$$
$$= P(A) + P(B) - P(A \cap B)$$

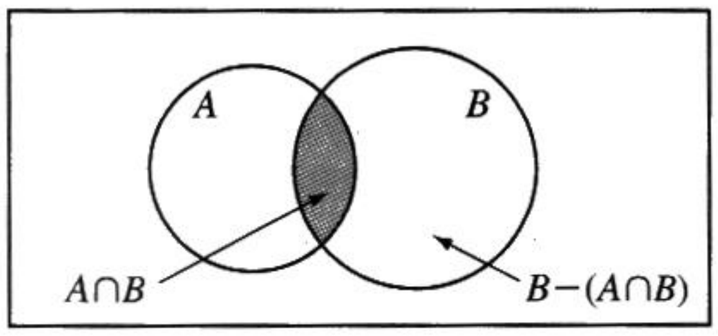


Fig. 1-7

CÁLCULO DE PROBABILIDADES

1.6. Se saca al azar una carta de una baraja de 52 cartas. Encuentre la probabilidad de que la carta sea (*a*) un as, (*b*) la jota de corazones, (*c*) un tres de tréboles o un seis de diamantes, (*d*) un corazón, (*e*) cualquier palo excepto corazones, (*f*) un diez o una pica, (*g*) ni un cuatro ni un trébol.

Para que sea más sencillo, usemos las letras C, P, D, T, para indicar corazones, picas, diamantes y tréboles, respectivamente, y 1, 2, . . ., 13 para indicar as, dos, . . ., rey. Entonces, $3 \cap C$ significa tres de corazones, mientras que $3 \cup C$ significa tres o corazones. Usemos el espacio muestral del ejemplo 1.1 (*b*), asignando probabilidades iguales de $1/52$ a cada punto muestral. Por ejemplo, $P(6 \cap \mathrm{T}) = 1/52$.

(*a*)
$$P(1) = P(1 \cap C \text{ o } 1 \cap P \text{ o } 1 \cap D \text{ o } 1 \cap T)$$
$$= P(1 \cap C) + P(1 \cap P) + P(1 \cap D) + P(1 \cap T)$$
$$= \frac{1}{52} + \frac{1}{52} + \frac{1}{52} + \frac{1}{52} = \frac{1}{13}$$

Esto también se hubiera podido obtener a partir del espacio muestral del problema 1.1 (*a*), en donde cada punto muestral, en especial el as, tiene probabilidad de $1/13$. También se hubiera podido llegar al mismo resultado por simple razonamiento; hay 13 números, de manera que cada uno tiene $1/13$ de probabilidad de aparecer cuando se saca una carta.

(*b*) $P(11 \cap C) = \dfrac{1}{52}$

(*c*) $P(3 \cap T \text{ o } 6 \cap D) = P(3 \cap T) + P(6 \cap D) = \dfrac{1}{52} + \dfrac{1}{52} = \dfrac{1}{26}$

(*d*) $P(C) = P(1 \cap C \text{ o } 2 \cap C \text{ o } \ldots 13 \cap C) = \dfrac{1}{52} + \dfrac{1}{52} + \cdots + \dfrac{1}{52} = \dfrac{13}{52} = \dfrac{1}{4}$

Esto también se hubiera podido obtener mediante la observación de que hay cuatro palos y cada uno tiene la misma probabilidad de salir: $1/4$

(*e*) $P(C') = 1 - P(C) = 1 - \dfrac{1}{4} = \dfrac{3}{4}$ usando la parte (*d*) y el teorema 1-4, página 6.

(*f*) Dado que 10 y S no son mutuamente excluyentes, tenemos, a partir del teorema 1-6,

$$P(10 \cup P) = P(10) + P(P) - P(10 \cap P) = \frac{1}{13} + \frac{1}{4} - \frac{1}{52} = \frac{4}{13}$$

(*g*) La probabilidad de ni cuatro ni trébol se puede describir por $P(4' \cap T')$. Pero $4' \cap T' = (4 \cup T)'$ Por consiguiente,

$$P(4' \cap T') = P[(4 \cup T)'] = 1 - P(4 \cup T)$$
$$= 1 - [P(4) + P(T) - P(4 \cap T)]$$
$$= 1 - \left[\frac{1}{13} + \frac{1}{4} - \frac{1}{52}\right] = \frac{9}{13}$$

También podemos obtener este resultado observando que el diagrama favorable para este evento es el complemento del evento resaltado en la figura 1-8. Dado que el complemento tiene $52 - 16 = 36$ puntos muestrales, y cada punto muestral tiene una probabilidad asignada de $1/52$, la probabilidad requerida es $36/52 = 9/13$.

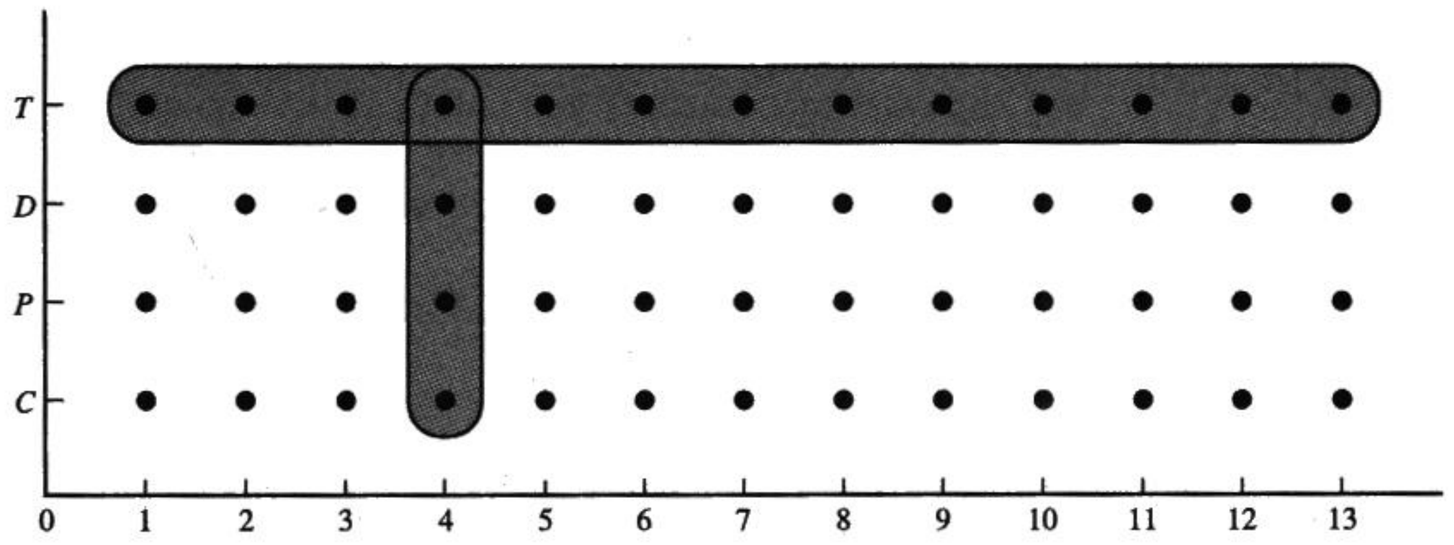


Fig. 1-8

1.7. Se saca una bola al azar de una caja que contiene 6 bolas rojas, 4 blancas y 5 azules. Determine la probabilidad de que la bola (*a*) sea roja, (*b*) blanca, (*c*) azul (*d*) no sea roja, (*e*) sea roja o blanca.

(*a*) **Método 1.**

Digamos que R, B y A describen los eventos de sacar una bola roja, blanca y azul, respectivamente. Entonces

$$P(R) = \frac{\text{maneras de escoger una bola roja}}{\text{maneras totales de escoger una bola}} = \frac{6}{6+4+5} = \frac{6}{15} = \frac{2}{5}$$

Método 2.

Nuestro espacio muestral consta de $6 + 4 + 5 = 15$ puntos muestrales. Entonces, si asignamos probabilidades iguales de $1/15$ a cada punto muestral, tenemos que $P(R) = 6/15 = 2/5$, puesto que hay 6 puntos muestrales que corresponden a "bola roja".

(*b*) $P(B) = \dfrac{4}{6+4+5} = \dfrac{4}{15}$

(*c*) $P(A) = \dfrac{5}{6+4+5} = \dfrac{5}{15} = \dfrac{1}{3}$

(*d*) $P(\text{no roja}) = P(R') = 1 - P(R) = 1 - \dfrac{2}{5} = \dfrac{3}{5}$ por la parte (*a*).

(*e*) **Método 1.**

$$P(\text{roja o blanca}) = P(R \cup B) = \frac{\text{maneras de escoger una bola roja o blanca}}{\text{maneras totales de escoger una bola}}$$

$$= \frac{6+4}{6+4+5} = \frac{10}{15} = \frac{2}{3}$$

Esto también puede resolverse usando el espacio muestral de la parte (*a*).

Método 2.

$$P(R \cup B) = P(A') = 1 - P(A) = 1 - \frac{1}{3} = \frac{2}{3} \text{ por la parte } (c).$$

Método 3.

Dado que los eventos R y B son mutuamente excluyentes, se obtiene de (*4*), página 6, que

$$P(R \cup B) = P(R) + P(B) = \frac{2}{5} + \frac{4}{15} = \frac{2}{3}$$

PROBABILIDAD CONDICIONAL Y EVENTOS INDEPENDIENTES

1.8. Un dado balanceado se lanza dos veces. Encuentre la probabilidad de obtener 4, 5 ó 6 en el primer lanzamiento y 1, 2, 3 ó 4 en el segundo.

Sea A_1 el evento "4, 5 ó 6 en el primer lanzamiento" y A_2 el evento "1, 2, 3 ó 4 en el segundo". Entonces, estamos buscando $P(A_1 \cap A_2)$.

Método 1.

$$P(A_1 \cap A_2) = P(A_1)P(A_2 \mid A_1) = P(A_1)P(A_2) = \left(\frac{3}{6}\right)\left(\frac{4}{6}\right) = \frac{1}{3}$$

Hemos utilizado aquí el hecho de que el resultado del segundo lanzamiento es *independiente* del primero, de manera que $P(A_2 \mid A_1) = P(A_2)$. También usamos $P(A_1) = 3/6$ (dado que 4, 5 ó 6 son 3 de los 6 resultados igualmente probables) y $P(A_2) = 4/6$ (dado que 1, 2, 3 ó 4 son 4 de los 6 resultados igualmente probables).

Método 2.

Cada una de las 6 maneras como puede caer un dado al primer lanzamiento puede asociarse con cada una de las 6 maneras como puede caer al segundo lanzamiento, un total de $6 \cdot 6 = 36$ maneras, todas igualmente probables.

Cada una de las 3 maneras como puede ocurrir A_1 se pueden asociar con cada una de las 4 maneras como puede ocurrir A_2, para obtener $3 \cdot 4 = 12$ maneras como pueden ocurrir A_1 y A_2. Entonces

$$P(A_1 \cap A_2) = \frac{12}{36} = \frac{1}{3}$$

Esto muestra directamente que A_1 y A_2 son independientes, dado que

$$P(A_1 \cap A_2) = \frac{1}{3} = \left(\frac{3}{6}\right)\left(\frac{4}{6}\right) = P(A_1)P(A_2)$$

1.9. Encuentre la probabilidad de no obtener 7 u 11 de la suma en ninguno de los dos lanzamientos de un par de dados balanceados.

El espacio muestral de cada lanzamiento se presenta en la figura 1-9. Por ejemplo, (5, 2) significa que 5 sale en el primer dado y 2 en el segundo. Puesto que los dados son balanceados y hay 36 puntos muestrales, asignamos a cada uno la probabilidad 1/36.

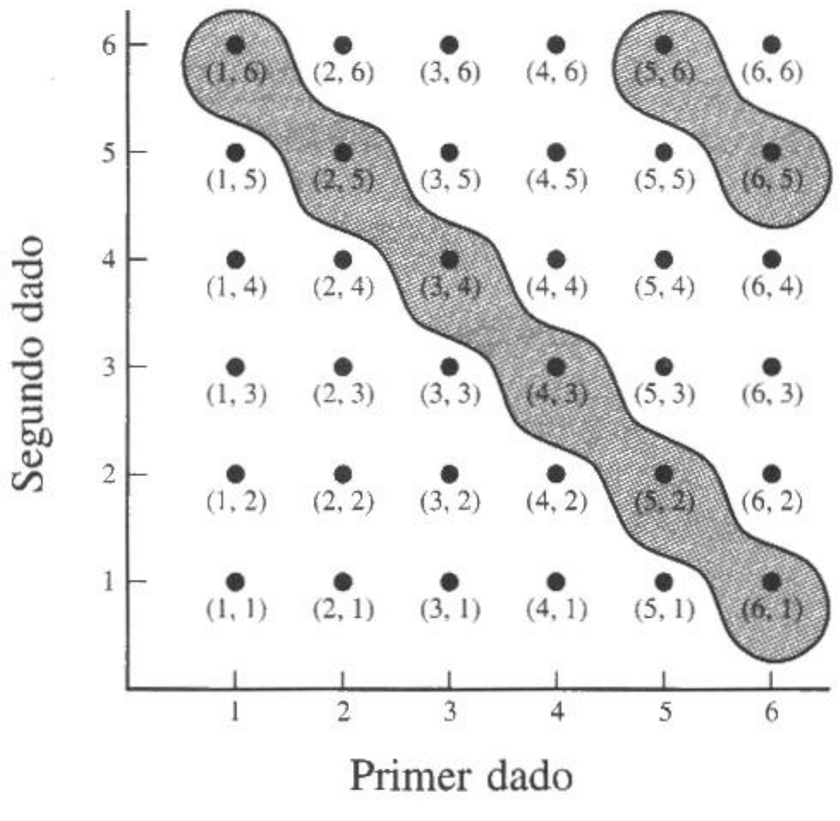


Fig. 1-9

Si A es el evento "7 u 11", entonces A se indica por la porción resaltada en la figura 1-9. Dado que se incluyen 8 puntos, tenemos que $P(A) = 8/36 = 2/9$. Se deduce que la probabilidad de no 7 u 11 está dada por

$$P(A') = 1 - P(A) = 1 - \frac{2}{9} = \frac{7}{9}$$

Utilizando los subíndices 1, 2 para designar el primer y segundo lanzamiento de los dados, vemos que la probabilidad de no 7 u 11 en cualquiera de los dos lanzamientos está dada por

$$P(A_1')P(A_2' \mid A_1') = P(A_1')P(A_2') = \left(\frac{7}{9}\right)\left(\frac{7}{9}\right) = \frac{49}{81},$$

teniendo en cuenta que los lanzamientos son independientes.

1.10. Se sacan dos cartas de un naipe bien barajado de 52 cartas. Encuentre la probabilidad de que ambas cartas sean ases si (*a*) hay remplazo, (*b*) no hay remplazo.

Método 1.

Sea A_1 = evento "as en el primer retiro" y A_2 = evento "as en el segundo retiro". Entonces estamos buscando $P(A_1 \cap A_2) = P(A_1)P(A_2 \mid A_1)$.

(*a*) Dado que para el primer retiro hay 4 ases en 52 cartas, $P(A_1) = 4/52$. Además, si se remplaza la carta antes de hacer el segundo retiro, entonces $P(A_2 \mid A_1) = 4/52$, puesto que también hay 4 ases en las 52 cartas. Entonces

$$P(A_1 \cap A_2) = P(A_1)P(A_2 \mid A_1) = \left(\frac{4}{52}\right)\left(\frac{4}{52}\right) = \frac{1}{169}$$

(*b*) Como en la parte (*a*), $P(A_1) = 4/52$. Sin embargo, si sale un as en el primer retiro, quedarán sólo 3 ases en las 51 cartas disponibles, de manera que $P(A_2 \mid A_1) = 3/51$. Entonces

$$P(A_1 \cap A_2) = P(A_1)P(A_2 \mid A_1) = \left(\frac{4}{52}\right)\left(\frac{3}{51}\right) = \frac{1}{221}$$

Método 2.

(*a*) La primera carta se puede sacar de cualquiera de 52 maneras, y dado que hay remplazo, la segunda carta también se puede sacar de cualquiera de 52 maneras. Entonces ambas cartas pueden sacarse en (52)(52) maneras, todas igualmente probables.

Aquí hay 4 maneras de escoger un as en el primer retiro y 4 en el segundo, de manera que el número de maneras de escoger ases en el primero y segundo retiros es (4)(4). Entonces la probabilidad requerida es

$$\frac{(4)(4)}{(52)(51)} = \frac{1}{169}$$

(*b*) La primera carta se puede retirar en cualquiera de 52 maneras, y dado que no hay remplazo, la segunda carta se puede retirar en cualquiera de 51 maneras. Entonces ambas cartas pueden retirarse en (52)(51) maneras, todas igualmente probables.

Aquí hay 4 maneras de escoger un as en el primer retiro y 3 maneras en el segundo retiro, de manera que el número de maneras de escoger ases en el primero y segundo retiro es (4)(3). Entonces la probabilidad requerida es

$$\frac{(4)(3)}{(52)(51)} = \frac{1}{221}$$

1.11. Sucesivamente se sacan tres bolas de la caja del problema 1.7. Encuentre la probabilidad de que se saquen en el orden: rojo, blanco y azul si cada bola (*a*) se remplaza, (*b*) no se remplaza.

Sea R_1 = evento "rojo en el primer retiro", B_2 = evento "blanco en el segundo retiro", A_3 = evento "azul en el tercer retiro". Requerimos $P(R_1 \cap B_2 \cap A_3)$.

(*a*) Si se reemplaza cada bola, entonces los eventos son independientes y

$$P(R_1 \cap B_2 \cap A_3) = P(R_1)P(B_2 \mid R_1)P(A_3 \mid R_2 \cap B_2)$$
$$= P(R_1)P(B_2)P(A_3)$$
$$= \left(\frac{6}{6+4+5}\right)\left(\frac{4}{6+4+5}\right)\left(\frac{5}{6+4+5}\right) = \frac{8}{225}$$

(*b*) Si no se remplaza cada bola, entonces los eventos son dependientes y

$$P(R_1 \cap B_2 \cap A_3) = P(R_1)P(B_2 \mid R_1)P(A_3 \mid R_1 \cap B_2)$$
$$= \left(\frac{6}{6+4+5}\right)\left(\frac{4}{5+4+5}\right)\left(\frac{5}{5+3+5}\right) = \frac{4}{91}$$

1.12. Encuentre la probabilidad de obtener al menos un 4 en dos lanzamientos de un dado balanceado.

Sea A_1 = evento "4 en el primer lanzamiento" y A_2 = evento "4 en el segundo lanzamiento". Entonces

$A_1 \cup A_2$ = evento "4 en el primer lanzamiento o 4 en el segundo o ambos"
= evento "al menos se obtiene un 4"

y requerimos $P(A_1 \cup A_2)$.

Método 1.

Los eventos A_1 y A_2 no son mutuamente excluyentes, pero son independientes. Por consiguiente, por (*10*) y (*21*),

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$
$$= P(A_1) + P(A_2) - P(A_1)P(A_2)$$
$$= \frac{1}{6} + \frac{1}{6} - \left(\frac{1}{6}\right)\left(\frac{1}{6}\right) = \frac{11}{36}$$

Método 2.

$$P(\text{al menos se obtiene un 4}) + P(\text{no se obtiene un 4}) = 1$$

Entonces
$$P(\text{al menos se obtiene un 4}) = 1 - P(\text{no se obtiene un 4})$$
$$= 1 - P(\text{no se obtiene un 4 en el primer lanzamiento ni en el segundo lanzamiento})$$
$$= 1 - P(A_1' \cap A_2') = 1 - P(A_1')P(A_2')$$
$$= 1 - \left(\frac{5}{6}\right)\left(\frac{5}{6}\right) = \frac{11}{36}$$

Método 3.

El número total de maneras igualmente probables como pueden caer ambos dados $= 6 \cdot 6 = 36$

También
Número de maneras como ocurre A_1 pero no $A_2 = 5$
Número de maneras como ocurre A_2 pero no $A_1 = 5$
Número de maneras como ocurren A_1 y $A_2 = 1$

Entonces el número de maneras como ocurre al menos uno de los eventos A_1 o $A_2 = 5 + 5 + 1 = 11$. Por consiguiente, $P(A_1 \cup A_2) = 11/36$.

1.13. Una bolsa contiene 4 bolas blancas y 2 negras; otra bolsa contiene 3 bolas blancas y 5 negras. Si se saca una bola de cada bolsa, encuentre la probabilidad de que (*a*) ambas sean blancas, (*b*) ambas sean negras, (*c*) una sea blanca y la otra negra.

Sea B_1 = evento "bola blanca de la primera bolsa", B_2 = evento "bola blanca de la segunda bolsa".

(*a*) $$P(B_1 \cap B_2) = P(B_1)P(B_2 \mid B_1) = P(B_1)P(B_2) = \left(\frac{4}{4+2}\right)\left(\frac{3}{3+5}\right) = \frac{1}{4}$$

(*b*) $$P(B_1' \cap B_2') = P(B_1')P(B_2' \mid B_1') = P(B_1')P(B_2') = \left(\frac{2}{4+2}\right)\left(\frac{5}{3+5}\right) = \frac{5}{24}$$

(*c*) La probabilidad requerida es

$$1 - P(B_1 \cap B_2) - P(B_1' \cap B_2') = 1 - \frac{1}{4} - \frac{5}{24} = \frac{13}{24}$$

1.14. Demuestre el teorema 1-10, página 9.

Demostraremos el teorema para el caso $n = 2$. Es fácil hacer extensiones para valores mayores de n.

Si el evento A debe resultar en uno de los dos eventos mutuamente excluyentes A_1, A_2, entonces

$$A = (A \cap A_1) \cup (A \cap A_2)$$

Pero $A \cap A_1$ y $A \cap A_2$ son mutuamente excluyentes dado que A_1 y A_2 lo son. Por lo tanto, a partir del axioma 3.

$$\begin{aligned} P(A) &= P(A \cap A_1) + P(A \cap A_2) \\ &= P(A_1)P(A \mid A_1) + P(A_2)P(A \mid A_2) \end{aligned}$$

Usando (*18*), página 8.

1.15. La caja *I* contiene 3 canicas rojas y 2 azules, mientas que la caja *II* contiene 2 canicas rojas y 8 azules. Se lanza una moneda balanceada. Si se obtiene cara, se saca una canica de la caja *I*; si se obtiene sello, se saca una canica de la caja *II*. Encuentre la probabilidad de que se saque una canica roja.

Sea R el evento "se escoge una canica roja" mientras que I y II describen los eventos de que se escojan las cajas I y II, respectivamente. Dado que se puede obtener una canica roja tanto en la caja I como en la II, podemos usar los resultados del problema 1.14 siendo $A = R$, $A_1 = I$, $A_2 = II$. Por lo tanto, la probabilidad de escoger una canica roja es

$$P(R) = P(I)P(R \mid I) + P(II)P(R \mid II) = \left(\frac{1}{2}\right)\left(\frac{3}{3+2}\right) + \left(\frac{1}{2}\right)\left(\frac{2}{2+8}\right) = \frac{2}{5}$$

TEOREMA DE BAYES

1.16. Demuestre el teorema de Bayes (teorema 1-11, página 9).

Dado que A se obtiene de uno de los eventos mutuamente excluyentes $A_1, A_2, \ldots, A_n$, tenemos del teorema 1-10 (problema 1.14),

$$P(A) = P(A_1)P(A \mid A_1) + \cdots + P(A_n)P(A \mid A_n) = \sum_{j=1}^{n} P(A_j)P(A \mid A_j)$$

Por lo tanto, $$P(A_k \mid A) = \frac{P(A_k \cap A)}{P(A)} = \frac{P(A_k)P(A \mid A_k)}{\sum_{j=1}^{n} P(A_j)P(A \mid A_j)}$$

1.17. Del problema 1.15 suponga que quien lanza la moneda no revela si obtiene cara o sello (de manera que tampoco se sabe de cuál caja sacó la canica), pero sí dice que sacó una canica roja. ¿Cuál es la probabilidad de que haya escogido la caja I (es decir, de que la moneda sea cara)?

Usemos la misma terminología que en el problema 1.15, es decir $A = R$, $A_1 = I$, $A_2 = II$. Busquemos la probabilidad de que la caja I haya sido escogida dado que sabemos que la persona sacó una canica roja. Usando el teorema de la regla de Bayes con $n = 2$, la probabilidad está dada por

$$P(I \mid R) = \frac{P(I)P(R \mid I)}{P(I)P(R \mid I) + P(II)P(R \mid II)} = \frac{\left(\frac{1}{2}\right)\left(\frac{3}{3+2}\right)}{\left(\frac{1}{2}\right)\left(\frac{3}{3+2}\right) + \left(\frac{1}{2}\right)\left(\frac{2}{2+8}\right)} = \frac{3}{4}$$

ANÁLISIS COMBINATORIO, CONTEO Y DIAGRAMAS DE ÁRBOL

1.18. Se va a conformar un comité de tres miembros, compuesto por un representante de los trabajadores, uno de la administración y uno del público en general. Si hay 3 posibles representantes de los trabajadores, 2 de la administración y 4 del público, determine cuántos comités diferentes pueden formarse usando (*a*) el principio fundamental del conteo y (*b*) un diagrama de árbol.

(*a*) Podemos escoger un representante de los trabajadores de 3 maneras diferentes, y luego un representante de la administración, de 2 maneras diferentes. Entonces hay $3 \cdot 2 = 6$ maneras diferentes de escoger un representante de los trabajadores y uno de la administración. Con cada una de esas maneras podemos escoger a un representante del público de 4 maneras distintas. Por lo tanto, el número de comités diferentes que se pueden formar es $3 \cdot 2 \cdot 4 = 24$.

(*b*) Designemos los 3 representantes de los trabajadores como L_1, L_2 y L_3; los representantes de la administración como M_1, M_2, y los representantes del público como P_1, P_2, P_3, P_4. Entonces, el diagrama de la figura 1-10 muestra que hay 24 comités diferentes en total. A partir de este diagrama de árbol podemos hacer una lista de todas las comisiones, por ejemplo, $L_1M_1P_1$, $L_1M_1P_2$, etc.

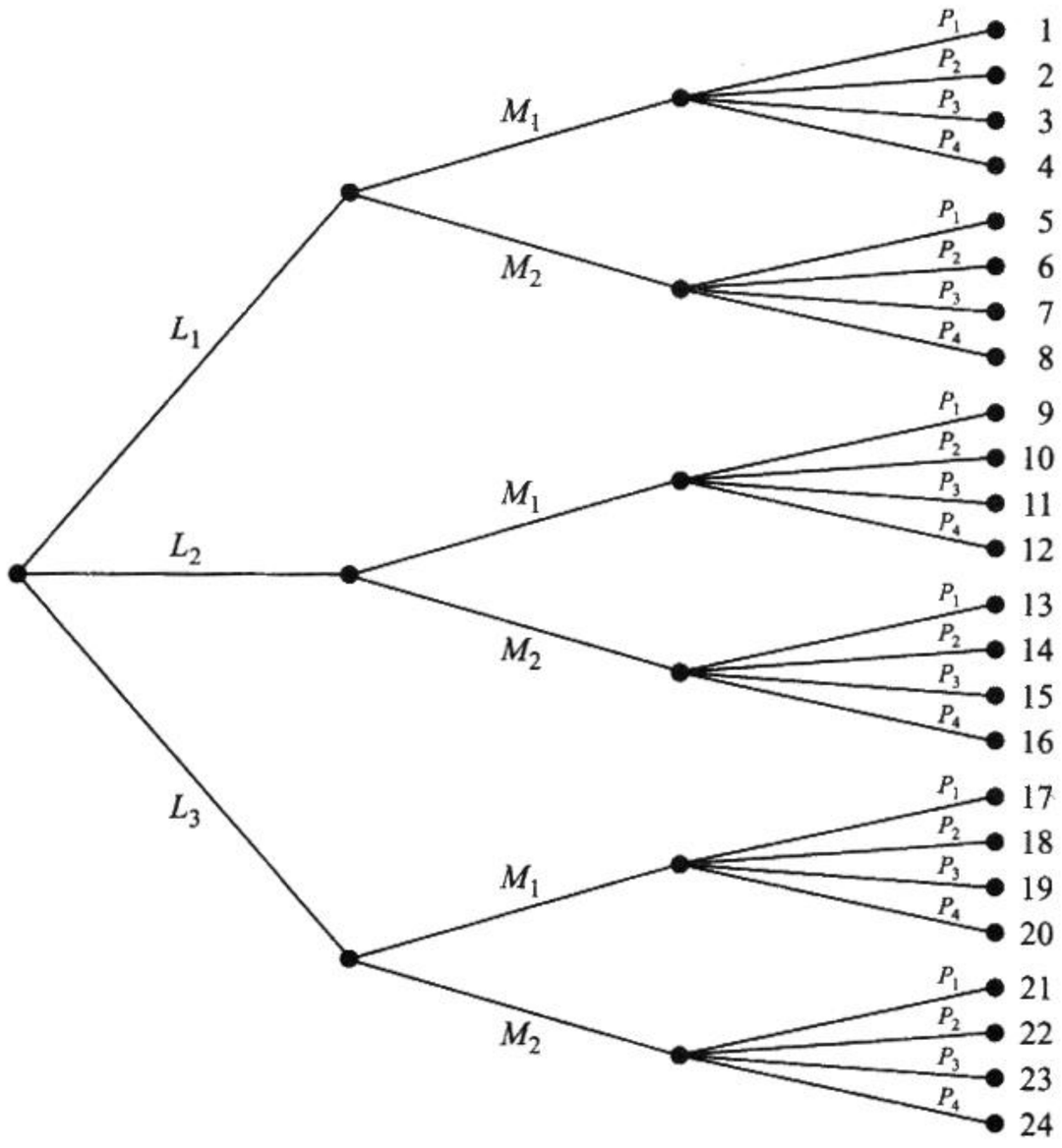


Fig. 1-10

PERMUTACIONES

1.19. ¿De cuántas maneras posibles se pueden ordenar 5 canicas de colores en una línea?

Debemos ordenar las 5 canicas en 5 posiciones así: — — — — —. Cualquiera de las canicas puede ocupar la primera posición, es decir, hay 5 maneras de llenar dicho lugar. Cuando se ha hecho esto, hay 4 maneras de llenar la segunda posición. Luego hay 3 maneras de llenar la tercera, 2 maneras de llenar la segunda y, por último, sólo 1 manera de llenar el último lugar. Por consiguiente:

$$\text{El número de ordenaciones de las 5 canicas en una línea} = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5! = 120$$

En general,

$$\text{El número de ordenaciones de } n \text{ objetos en una línea} = n(n-1)(n-2)\cdots 1 = n!$$

Esto se conoce también como *el número de permutaciones de n objetos diferentes tomados de n a la vez* y se denota como ${}_nP_n$.

1.20. ¿De cuántas maneras posibles se pueden sentar 10 personas en una banca si solamente hay 4 puestos disponibles?

El primer puesto puede ocuparse de cualquiera de 10 maneras, luego el segundo puede ocuparse de 9 maneras, el tercero de 8 maneras diferentes y el cuarto de 7. Por lo tanto,

$$\text{El número de ordenaciones de 10 personas tomadas de 4 a la vez} = 10 \cdot 9 \cdot 8 \cdot 7 = 5040$$

En general,

$$\text{El número de ordenaciones de } n \text{ objetos diferentes tomados de } r \text{ a la vez} = n(n-1)\cdots(n-\mathrm{r}+1)$$

Esto también se conoce como *el número de permutaciones de n objetos diferentes tomados de r a la vez* y se denota por ${}_nP_r$. Observe que cuando $r = n$, ${}_nP_n = n!$ como en el problema 1.19.

1.21. Calcule (*a*) ${}_8P_3$, (*b*) ${}_6P_4$, (*c*) ${}_{15}P_1$, (*d*) ${}_3P_3$.

(*a*) ${}_8P_3 = 8 \cdot 7 \cdot 6 = 336$ (*b*) ${}_6P_4 = 6 \cdot 5 \cdot 4 \cdot 3 = 360$ (*c*) ${}_{15}P_1 = 15$ (*d*) ${}_3P_3 = 3 \cdot 2 \cdot 1 = 6$

1.22. Se necesita sentar 5 hombres y 4 mujeres en fila, de manera que las mujeres ocupen los lugares pares. ¿De cuántas maneras pueden sentarse?

Los hombres se pueden sentar de ${}_5P_5$ maneras, y las mujeres de ${}_4P_4$. Cada arreglo de los hombres puede asociarse con cada arreglo de las mujeres. Entonces,

$$\text{Número de arreglos} = {}_5P_5 \cdot {}_4P_4 = 5!4! = (120)(24) = 2880$$

1.23. ¿Cuántos números de 4 dígitos pueden formarse con los 10 números 0, 1, 2, 3, . . ., 9 si (*a*) se permiten repeticiones, (*b*) no se permiten repeticiones, (*c*) el último número debe ser cero y no se permiten repeticiones?

(*a*) El primer número puede ser cualesquiera de los 9 dígitos (dado que el cero no es permitido). El segundo, tercero y cuarto números pueden ser cualquiera de los 10. Entonces $9 \cdot 10 \cdot 10 \cdot 10 = 9000$ números que se pueden formar.

(*b*) El primer número puede ser cualquiera de los 9 (cualquiera menos el 0).

El segundo número puede ser cualquiera de los 9 (cualquiera menos el que se usó para el primer dígito).

El tercer número puede ser cualquiera de los 8 (cualquiera menos los que se usaron para los dos primeros dígitos).

El cuarto número puede ser cualquiera de los 7 (cualquiera menos los que se usaron para los tres primeros dígitos).

Entonces $9 \cdot 9 \cdot 8 \cdot 7 = 4536$ números que se pueden formar.

Otro método.

El primer dígito puede ser cualquiera de los 9 números, y los tres restantes se pueden escoger de ${}_9P_3$ maneras. Entonces, $9 \cdot 9 \cdot 8 \cdot 7 = 4536$ números que se pueden formar.

(*c*) El primer dígito se puede escoger de 9 maneras, el segundo de 8 maneras y el tercero de 7 maneras. Entonces $9 \cdot 8 \cdot 7 = 504$ números que se pueden formar.

Otro método.

El primer dígito se puede escoger de 9 maneras, y los siguientes dos se pueden escoger de ${}_8P_2$ maneras. Entonces, $9 \cdot {}_8P_2 = 9 \cdot 8 \cdot 7 = 504$ números que se pueden formar.

1.24. En una biblioteca se quieren acomodar cuatro libros diferentes de matemáticas, seis libros distintos de física y dos libros diferentes de química. ¿Cuántos ordenamientos diferentes son posibles si (*a*) los libros de cada materia deben quedar juntos, (*b*) sólo los libros de matemáticas deben quedar juntos?

(*a*) Los libros de matemáticas se pueden ordenar entre ellos de ${}_4P_4 = 4!$ maneras, los de física de ${}_6P_6 = 6!$ maneras, los de química de ${}_2P_2 = 2!$ maneras y los tres grupos de libros de ${}_3P_3 = 3!$ maneras. Por consiguiente,

$$\text{Número de ordenamientos} = 4!6!2!3! = 207{,}360$$

(*b*) Consideremos los cuatro libros de matemáticas como si fueran uno sólo. Entonces, nos quedan 9 libros que debemos acomodar de ${}_9P_9 = 9!$ maneras. En todos estos arreglos, los libros de matemáticas están juntos. Pero los de matemáticas se pueden arreglar entre ellos de ${}_4P_4$ maneras. Por lo tanto,

$$\text{Número de arreglos} = 9!4! = 8{,}709{,}120$$

1.25. En una línea están acomodadas cinco canicas rojas, dos blancas y tres azules. Si las canicas del mismo color no pueden diferenciarse entre sí, ¿cuántos arreglos diferentes se pueden hacer?

Supongamos que hay N arreglos diferentes. Multiplicando N por el número de maneras de arreglar (*a*) las cinco canicas rojas entre ellas, (*b*) las dos canicas blancas entre ellas y (*c*) las tres canicas azules entre ellas (es decir, multiplicando N por 5!2!3!), obtenemos el número de maneras de arreglar las 10 canicas si se distinguen entre sí, es decir, 10!

Entonces $$(5!2!3!)N = 10! \quad \text{y} \quad N = 10!/(5!2!3!)$$

En general, el número de arreglos diferentes de n objetos, de los cuales n_1 son iguales, n_2 son iguales, . . ., n_k son iguales, es $\dfrac{n!}{n_1!\,n_2!\cdots n_k!}$ donde $n_1 + n_2 + \cdots + n_k = n$.

1.26. ¿De cuántas maneras posibles se pueden sentar 7 personas alrededor de una mesa redonda si (*a*) se pueden sentar en cualquier lugar, (*b*) 2 personas en particular no se pueden sentar juntas?

(*a*) Empecemos con una persona sentada en cualquier lugar. Entonces, las seis personas restantes se pueden sentar de $6! = 720$ maneras, las cuales son el número total de maneras como se pueden sentar 7 personas en círculo.

(b) Consideremos a las dos personas en particular como si fueran una sola. Entonces, hay 6 personas en total, que pueden acomodarse de 5! maneras. Pero, las dos personas que consideramos como si fueran una pueden, a su vez, ordenarse de 2! maneras. Por consiguiente, el número de maneras de organizar a 7 personas alrededor de una mesa redonda con 2 personas en particular sentadas juntas $= 5!2! = 240$.

Entonces usando (*a*), el número total de maneras como pueden sentarse 7 personas alrededor de una mesa redonda de manera que 2 personas en particular no se sienten juntas $= 730 - 240 = 480$ maneras.

COMBINACIONES

1.27. ¿De cuántas maneras se pueden dividir 10 objetos en dos grupos que contengan 4 y 6 objetos respectivamente?

Esto es lo mismo que el número de ordenaciones de 10 objetos, de los cuales 4 son semejantes entre sí y los otros 6 también. Según el problema 1.25, esto es $\dfrac{10!}{4!\,6!} = \dfrac{10 \cdot 9 \cdot 8 \cdot 7}{4!} = 210$.

El problema equivale a encontrar el número de selecciones de 4 de 10 objetos (o 6 de 10 objetos), siendo irrelevante el orden de la selección. En general, el número de selecciones de r de n objetos, llamado *el número de combinaciones de n objetos tomados r a la vez*, se describe por ${}_nC_r$ o $\binom{n}{r}$ y está dado por

$$ {}_nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{n(n-1)\cdots(n-r+1)}{r!} = \frac{{}_nP_r}{r!} $$

1.28. Calcule (*a*) ${}_7C_4$, (*b*) ${}_6C_5$, (*c*) ${}_4C_4$.

(*a*) $ {}_7C_4 = \dfrac{7!}{4!\,3!} = \dfrac{7 \cdot 6 \cdot 5 \cdot 4}{4!} = \dfrac{7 \cdot 6 \cdot 5}{3 \cdot 2 \cdot 1} = 35. $

(*b*) $ {}_6C_5 = \dfrac{6!}{5!\,1!} = \dfrac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2}{5!} = 6, \qquad \text{o} \qquad {}_6C_5 = {}_6C_1 = 6. $

(*c*) ${}_4C_4$ es el número de selecciones de 4 objetos tomados de 4 a la vez, y sólo hay una de dichas selecciones. Entonces ${}_4C_4 = 1$. Observe formalmente que

$$ {}_4C_4 = \frac{4!}{4!\,0!} = 1 \text{ si } \textit{definimos } 0! = 1. $$

1.29. ¿De cuántas maneras se puede formar un comité de 5 personas a partir de un grupo de 9?

$$ \binom{9}{5} = {}_9C_5 = \frac{9!}{5!\,4!} = \frac{9 \cdot 8 \cdot 7 \cdot 6 \cdot 5}{5!} = 126 $$

1.30. Se va a formar un comité de 2 matemáticos y 3 físicos a partir de 5 matemáticos y 7 físicos. ¿De cuántas maneras se puede hacer si (*a*) se puede incluir cualquier matemático y cualquier físico, (*b*) un físico en particular debe estar en el comité y (*c*) dos matemáticos en particular no pueden pertenecer al comité?

(*a*) Se pueden seleccionar 2 matemáticos de 5 de ${}_5C_2$ maneras.
Se pueden seleccionar 3 físicos de 7 de ${}_7C_3$ maneras.

$$ \text{Número total de selecciones posibles} = {}_5C_2 \cdot {}_7C_3 = 10 \cdot 35 = 350 $$

(*b*) Se pueden seleccionar 2 matemáticos de 5 de ${}_5C_2$ maneras.
Se pueden seleccionar 2 físicos de 6 de ${}_7C_2$ maneras.

$$ \text{Número total de selecciones posibles} = {}_5C_2 \cdot {}_6C_2 = 10 \cdot 15 = 150 $$

(*c*) Se pueden seleccionar 2 matemáticos de 3 de ${}_3C_2$ maneras.
Se pueden seleccionar 3 físicos de 7 de ${}_7C_3$ maneras.

$$ \text{Número total de selecciones posibles} = {}_3C_2 \cdot {}_7C_3 = 3 \cdot 35 = 105 $$

1.31. ¿Cuántas ensaladas diferentes se pueden hacer con lechuga, escarola, endibia, berro y achicoria?

Cada vegetal puede tratarse de 2 maneras, según se escoja o se rechace un vegetal. Puesto que cada 2 formas de tratar a un vegetal está asociada con las 2 formas de tratar a cada uno de los otros vegetales, el número de maneras de tratar a los 5 vegetales es 2^5. Pero dentro de las 2^5 maneras se incluye el caso de que no se seleccione ningún vegetal. Por lo tanto,

$$\text{Número de ensaladas} = 2^5 - 1 = 31$$

Otro método.

Se puede seleccionar 1 de los 5 vegetales, 2 de los 5, . . ., 5 de los 5. Entonces el número de ensaladas es

$$ {}_5C_1 + {}_5C_2 + {}_5C_3 + {}_5C_4 + {}_5C_5 = 5 + 10 + 10 + 5 + 1 = 31$$

En general, para cualquier número entero positivo n, ${}_nC_1 + {}_nC_2 + {}_nC_3 + {}_nC_4 + {}_nC_n = 2^n - 1$

1.32. ¿Cuántas palabras de 4 consonantes diferentes y 3 vocales distintas se pueden formar a partir de 7 consonantes y 5 vocales? No es necesario que las palabras tengan significado.

Se pueden seleccionar las 4 consonantes diferentes de ${}_7C_4$ maneras y las 3 vocales diferentes de ${}_5C_3$ maneras. Además, las 7 letras resultantes (4 consonantes y 3 vocales) pueden ordenarse entre sí de ${}_7P_7 = 7!$ maneras. Entonces

$$\text{Número de palabras} = {}_7C_4 \cdot {}_5C_3 \cdot 7! = 35 \cdot 10 \cdot 5040 = 1{,}764{,}000$$

COEFICIENTES BINOMIALES

1.33. Demuestre que $\binom{n}{r} = \binom{n-1}{r} + \binom{n-1}{r-1}$.

Tenemos que

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{n(n-1)!}{r!(n-r)!} = \frac{(n-r+r)(n-1)!}{r!(n-r)!}$$

$$= \frac{(n-r)(n-1)!}{r!(n-r)!} + \frac{r(n-1)!}{r!(n-r)!}$$

$$= \frac{(n-1)!}{r!(n-r-1)!} + \frac{(n-1)!}{(r-1)!(n-r)!}$$

$$= \binom{n-1}{r} + \binom{n-1}{r-1}$$

El resultado tiene la siguiente aplicación interesante. Si escribimos los coeficientes de la expansión binomial de $(x + y)^n$ para $n = 0, 1, 2, \ldots$, obtenemos el siguiente ordenamiento, llamado el *triángulo de Pascal*.

$n = 0$ 1
$n = 1$ 1 1
$n = 2$ 1 2 1
$n = 3$ 1 3 3 1
$n = 4$ 1 4 6 4 1
$n = 5$ 1 5 10 10 5 1
$n = 6$ 1 6 15 20 15 6 1
etc.

Se puede obtener un resultado en cualquier renglón sumando las dos entradas que se encuentran inmediatamente a la izquierda y derecha del renglón precedente. Por consiguiente, $10 = 4 + 6$, $15 = 10 + 5$, etc.

1.34. Encuentre el término constante en la expansión de $\left(x^2+\frac{1}{x}\right)^{12}$.

Según el teorema del binomio,

$$\left(x^2+\frac{1}{x}\right)^{12}=\sum_{k=0}^{12}\binom{12}{k}(x^2)^k\left(\frac{1}{x}\right)^{12-k}=\sum_{k=0}^{12}\binom{12}{k}x^{3k-12}$$

El término constante corresponde a aquél para el cual $3k-12=0$, es decir, $k=4$, y por consiguiente está dado por

$$\binom{12}{4}=\frac{12\cdot 11\cdot 10\cdot 9}{4\cdot 3\cdot 2\cdot 1}=495$$

PROBABILIDAD USANDO EL ANÁLISIS COMBINATORIO

1.35. Una caja contiene 8 bolas rojas, 3 blancas y 9 azules. Si se sacan 3 bolas al azar y sin remplazo, determine la probabilidad de que (*a*) las 3 sean rojas, (*b*) las 3 sean blancas, (*c*) 2 sean rojas y 1 blanca, (*d*) al menos 1 sea blanca, (*e*) se saque una de cada color, (*f*) se saquen en el siguiente orden: roja, blanca, azul.

(*a*) **Método 1.**

Sean R_1, R_2, R_3 los eventos "bola roja en la primera sacada", "bola roja en la segunda sacada", "bola roja en la tercera sacada", respectivamente. Entonces, $R_1\cap R_2\cap R_3$ denota el evento "todas las bolas sacadas son rojas". Tenemos entonces

$$P(R_1\cap R_2\cap R_3)=P(R_1)P(R_2\mid R_1)P(R_3\mid R_1\cap R_2)$$
$$=\left(\frac{8}{20}\right)\left(\frac{7}{19}\right)\left(\frac{6}{18}\right)=\frac{14}{285}$$

Método 2.

$$\text{Probabilidad requerida}=\frac{\text{número de selecciones de 3 bolas rojas entre 8}}{\text{número de selecciones de 3 bolas entre 20}}=\frac{{}_8C_3}{{}_{20}C_3}=\frac{14}{285}$$

(*b*) Usando el segundo método que se indica en la parte (*a*),

$$P(\text{las 3 bolas son blancas})=\frac{{}_3C_3}{{}_{20}C_3}=\frac{1}{1140}$$

También se puede usar el primer método indicado en la parte (*a*).

(*c*) $P(2 \text{ son rojas y } 1 \text{ es blanca})$

$$=\frac{(\text{selecciones de 2 bolas rojas entre 8})(\text{selecciones de 1 bola blanca entre 3})}{\text{número de selecciones de 3 bolas entre 20}}$$
$$=\frac{({}_8C_2)({}_3C_1)}{{}_{20}C_3}=\frac{7}{95}$$

(*d*) $P(\text{ninguna bola es blanca})=={}_{17}C_3/{}_{20}C_3=\frac{34}{57}$. Entonces

$$P(\text{al menos 1 bola es blanca})=1-\frac{34}{57}=\frac{23}{57}$$

(*e*) $P(\text{se saca una bola de cada color})=\dfrac{({}_8C_1)({}_3C_1)({}_9C_1)}{{}_{20}C_3}=\dfrac{18}{95}$

$$(f)\ P(\text{se sacan en el orden rojo, blanco, azul}) = \frac{1}{3!}\,P(\text{se saca una de cada color})$$

$$= \frac{1}{6}\left(\frac{18}{95}\right) = \frac{3}{95},\ \text{usando } (e)$$

Otro método.

$$P(R_1 \cap B_2 \cap A_3) = P(R_1)P(B_2 \mid R_1)P(A_3 \mid R_1 \cap B_2)$$

$$= \left(\frac{8}{20}\right)\left(\frac{3}{19}\right)\left(\frac{9}{18}\right) = \frac{3}{95}$$

1.36. En un juego de póker se sacan 5 cartas de un naipe de 52 cartas bien barajadas. Encuentre la probabilidad de (*a*) 4 ases, (*b*) 4 ases y 1 rey, (*c*) 3 diez y 2 jotas, (*d*) nueve, diez, jota, reina, rey en cualquier orden, (*e*) 3 cartas son de un palo y 2 de otro, (*f*) se saca al menos un as.

$$(a)\ P(4 \text{ ases}) = \frac{({}_4C_4)({}_{48}C_1)}{{}_{52}C_5} = \frac{1}{54{,}145}.$$

$$(b)\ P(4 \text{ ases y } 1 \text{ rey}) = \frac{({}_4C_4)({}_4C_1)}{{}_{52}C_5} = \frac{1}{649{,}740}.$$

$$(c)\ P(3 \text{ diez y } 2 \text{ jotas}) = \frac{({}_4C_3)({}_4C_2)}{{}_{52}C_5} = \frac{1}{108{,}290}.$$

$$(d)\ P(\text{nueve, diez, jota, reina, rey en cualquier orden}) = \frac{({}_4C_1)({}_4C_1)({}_4C_1)({}_4C_1)({}_4C_1)}{{}_{52}C_5} = \frac{64}{162{,}435}.$$

$$(e)\ P(3 \text{ cartas de cualquier palo, } 2 \text{ de otro palo}) = \frac{(4 \cdot {}_{13}C_3)(3 \cdot {}_{13}C_2)}{{}_{52}C_5} = \frac{429}{4165},$$

dado que hay 4 maneras de escoger el primer palo y 3 de escoger el segundo.

$$(f)\ P(\text{no as}) = \frac{{}_{48}C_5}{{}_{52}C_5} = \frac{35{,}673}{54{,}145}. \qquad \text{Entonces } P(\text{al menos un as}) = 1 - \frac{35{,}673}{54{,}145} = \frac{18{,}472}{54{,}145}.$$

1.37. Determine la probabilidad de obtener 3 seis al lanzar 5 veces un dado balanceado.

Los lanzamientos del dado se representarán con 5 espacios — — — — —. En cada espacio tendremos el evento 6 o no 6 (6′). Por ejemplo, tres 6 y dos no 6 pueden ocurrir como 666′66′ o 66′66′, etc.

Ahora, la probabilidad de que ocurra 666′66′ es

$$P(666'66') = P(6)P(6)P(6')P(6)P(6') = \frac{1}{6}\cdot\frac{1}{6}\cdot\frac{5}{6}\cdot\frac{1}{6}\cdot\frac{5}{6} = \left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2$$

dado que suponemos independencia. Igualmente,

$$P = \left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2$$

para los demás resultados en los que ocurren tres 6 y dos no 6. Pero hay ${}_5C_3 = 10$ de estos eventos, y son mutuamente excluyentes. Por consiguiente, la probabilidad es

$$P(666'66' \text{ o } 66'66'6 \text{ o } \ldots) = {}_5C_3\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2 = \frac{5!}{3!\,2!}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2 = \frac{125}{3888}$$

En general, si $p = P(A)$ y $q = 1 - p = P(A')$, entonces usando el mismo razonamiento que se utilizó anteriormente, la probabilidad de obtener exactamente x veces A en n ensayos independientes es

$$ {}_nC_x p^x q^{n-x} = \binom{n}{x} p^x q^{n-x} $$

1.38. Una biblioteca tiene 6 libros de matemáticas y 4 de física. Encuentre la probabilidad de que 3 libros de matemáticas en particular estén juntos.

Todos los libros pueden ordenarse entre sí de ${}_{10}P_{10} = 10!$ maneras. Supongamos que 3 libros determinados de matemáticas se remplazan por 1. Entonces tenemos un total de 8 libros que se pueden ordenar entre sí de ${}_8P_8 = 8!$ maneras. Pero los 3 libros de matemáticas en sí mismos se pueden ordenar de ${}_3P_3 = 3!$ maneras. Así, la probabilidad está dada por

$$ \frac{8!\,3!}{10!} = \frac{1}{15} $$

PROBLEMAS DIVERSOS

1.39. A y B juegan 12 partidas de ajedrez de las cuales A gana 6, B gana 4, y 2 terminan en empate. Ellos acuerdan jugar un torneo de 3 partidas. Encuentre la probabilidad de (*a*) A gana los 3 juegos, (*b*) 2 juegos finalizan en empate, (*c*) A y B ganan alternadamente y (*d*) B gana al menos 1 juego.

Llamemos A_1, A_2, A_3 a los eventos "A gana" el primero, segundo y tercer juego respectivamente; llamemos B_1, B_2, B_3 a los eventos "B gana" el primero, segundo y tercer juego respectivamente. Con base en su desempeño anterior (probabilidad empírica), supongamos que

$$ P(\text{A gana un juego}) = \frac{6}{12} = \frac{1}{2}, \quad P(\text{B gana un juego}) = \frac{4}{12} = \frac{1}{3} $$

(*a*) $$ P(A \text{ gana los 3 juegos}) = P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3) = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{8} $$

Suponiendo que el resultado de cada juego sea independiente del resultado de los otros juegos (esta suposición no se justificaría si cada jugador estuviera *psicológicamente influenciado* por la ganancia o pérdida del otro).

(*b*) En cualquier juego la probabilidad de no empate (es decir, A o B gana) es $q = \frac{1}{2} + \frac{1}{3} = \frac{5}{6}$ y la probabilidad de empate es $p = 1 - q = \frac{1}{6}$. Entonces la probabilidad de 2 empates en 3 ensayos es (*véase* problema 1.37)

$$ \binom{3}{2} p^2 q^{3-2} = 3\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right) = \frac{5}{72} $$

(*c*)
$$ \begin{aligned} P(\text{A y B ganan alternadamente}) &= P(\text{A gana, luego B gana, luego A gana} \\ &\quad o \text{ B gana, luego A gana, luego B gana}) \\ &= P(A_1 \cap B_2 \cap A_3) + P(B_1 \cap A_2 \cap B_3) \\ &= P(A_1)P(B_2)P(A_3) + P(B_1)P(A_2)P(B_3) \\ &= \left(\frac{1}{2}\right)\left(\frac{1}{3}\right)\left(\frac{1}{2}\right) + \left(\frac{1}{3}\right)\left(\frac{1}{2}\right)\left(\frac{1}{3}\right) = \frac{5}{36} \end{aligned} $$

(*d*)
$$ \begin{aligned} P(\text{B gana al menos un juego}) &= 1 - P(\text{B no gana}) \\ &= 1 - P(B_1' \cap B_2' \cap B_3') \\ &= 1 - P(B_1')P(B_2')P(B_3') \\ &= 1 - \left(\frac{2}{3}\right)\left(\frac{2}{3}\right)\left(\frac{2}{3}\right) = \frac{19}{27} \end{aligned} $$

1.40. A y B juegan una partida en la que alternadamente lanzan un par de dados. Gana el primero que obtenga un total de 7. Encuentre la probabilidad de (*a*) el primero que lanza gana el juego y (*b*) gana el que lanza en segundo lugar.

(*a*) La probabilidad de obtener 7 en el lanzamiento de un par de dados, suponiendo que éste es balanceado, es $\frac{1}{6}$, como se vio en el problema 1.9 y en la figura 1-9. Si suponemos que A lanza en primer lugar, entonces A ganará en cualquiera de los siguientes casos mutuamente excluyentes, que indican probabilidades asociadas:

(1) A gana en el primer lanzamiento. Probabilidad $= \frac{1}{6}$.

(2) A pierde en el primer lanzamiento, luego pierde B, luego gana A. Probabilidad $= (\frac{5}{6})(\frac{5}{6})(\frac{1}{6})$.

(3) A pierde en el primer lanzamiento, B pierde, A pierde, B pierde, A gana. Probabilidad $= (\frac{5}{6})(\frac{5}{6})(\frac{5}{6})(\frac{5}{6})(\frac{1}{6})$.

.....

Entonces la probabilidad de que A gane es

$$\left(\frac{1}{6}\right)+\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{1}{6}\right)+\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{1}{6}\right)+\cdots$$

$$=\frac{1}{6}\left[1+\left(\frac{5}{6}\right)^2+\left(\frac{5}{6}\right)^4+\cdots\right]=\frac{1/6}{1-(5/6)^2}=\frac{6}{11}$$

donde hemos usado el resultado del apéndice A con $x = (5/6)^2$.

(*b*) Igualmente la probabilidad de que B gane el juego es

$$\left(\frac{5}{6}\right)\left(\frac{1}{6}\right)+\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{1}{6}\right)+\cdots=\left(\frac{5}{6}\right)\left(\frac{1}{6}\right)\left[1+\left(\frac{5}{6}\right)^2+\left(\frac{5}{6}\right)^4+\cdots\right]$$

$$=\frac{5/36}{1-(5/6)^2}=\frac{5}{11}$$

Por consiguiente, apostaremos 6 a 5 que el que lance primero gana. Observe que, dado que

$$\frac{6}{11}+\frac{5}{11}=1$$

la probabilidad de un empate es cero. Esto no sería cierto si el juego fuera limitado. *Véase* problema 1.100.

1.41. Una máquina produce un total de 12,000 tornillos cada día, de los cuales el 3% en promedio es defectuoso. Encuentre la probabilidad de que de 600 tornillos tomados al azar, haya 12 defectuosos.

De los 12,000 tornillos el 3% es defectuoso, es decir, 360 tornillos son defectuosos y 11,640 no. Entonces:

$$\text{Probabilidad requerida} = \frac{{}_{360}C_{12}\,{}_{11,640}C_{588}}{{}_{12,000}C_{600}}$$

1.42. Una caja contiene 5 canicas rojas y 4 blancas. Se sacan, una tras otra, dos canicas de la caja sin remplazo, y se observa que la segunda canica es blanca. ¿Cuál es la probabilidad de que la primera también sea blanca?

Método 1.

Si W_1, W_2 son los eventos "blanca en el primer retiro", "blanca en el segundo retiro"; respectivamente, estamos buscando $P(W_1 \mid W_2)$. Esto está dado por

$$P(W_1 \mid W_2)=\frac{P(W_1 \cap W_2)}{P(W_2)}=\frac{(4/9)(3/8)}{4/9}=\frac{3}{8}$$

Método 2.

Dado que se sabe que la segunda canica es blanca, sólo hay tres maneras de que de las 8 canicas restantes la primera también sea blanca, así que la probabilidad es 3/8.

1.43. Las probabilidades de que un marido y su esposa estén vivos durante 20 años a partir de ahora está dada por 0.8 y 0.9, respectivamente. Encuentre la probabilidad de que en 20 años estén vivos (*a*) ambos, (*b*) ninguno, (*c*) al menos uno.

Sean H, B los eventos de que el marido y la esposa estén vivos dentro de 20 años, respectivamente. Entonces $P(H) = 0.8$, $P(W) = 0.9$. Supongamos que H y W son eventos independientes, lo cual puede o no ser razonable.

(*a*) $P(\text{ambos estén vivos}) = P(H \cap W) = P(H)\,P(W) = (0.8)(0.9) = 0.72$

(*b*) $P(\text{ninguno esté vivo}) = P(H' \cap W') = P(H')\,P(W') = (0.2)(0.1) = 0.02$

(*c*) $P(\text{al menos uno esté vivo}) = 1 - P(\text{ninguno esté vivo}) = 1 - 0.02 = 0.98$

1.44. Una secretaria ineficiente coloca al azar n cartas diferentes en n sobres marcados con destinatarios distintos. Encuentre la probabilidad de que al menos una de las cartas llegue al destinatario correcto.

Sean $A_1, A_2, \ldots, A_n$ denotan los eventos en que la primera, segunda, . . ., n-ésima carta esté en el sobre correcto. Entonces, el evento de que al menos una carta esté en el sobre correcto es $A_1 \cup A_2 \cup \cdots \cup A_n$, y queremos encontrar $P(A_1 \cup A_2 \cup \cdots \cup A_n)$. A partir de la generalización de los resultados (*10*) y (*11*), página 5, 6 y 7, tenemos

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum P(A_k) - \sum P(A_j \cap A_k) + \sum P(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n-1} P(A_1 \cap A_2 \cap \cdots \cap A_n) \tag{1}$$

Donde $\sum P(A_k)$ es la suma de las probabilidades de A_k desde 1 hasta n, $\sum P(A_j \cap A_k)$ es la suma de las probabilidades de $A_j \cap A_k$ con j y k de 1 hasta n y $k > j$, etc. Tenemos, por ejemplo, lo siguiente:

$$P(A_1) = \frac{1}{n} \quad \text{y similarmente} \quad P(A_k) = \frac{1}{n} \tag{2}$$

ya que de los n sobres sólo 1 tendrá la dirección correcta. También

$$P(A_1 \cap A_2) = P(A_1)P(A_2 \mid A_1) = \left(\frac{1}{n}\right)\left(\frac{1}{n-1}\right) \tag{3}$$

ya que si la primera carta está en el sobre correcto, entonces sólo 1 de los $n - 1$ sobres restantes será adecuado. De igual manera encontramos

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) = \left(\frac{1}{n}\right)\left(\frac{1}{n-1}\right)\left(\frac{1}{n-2}\right) \tag{4}$$

etc., y finalmente,

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = \left(\frac{1}{n}\right)\left(\frac{1}{n-1}\right)\cdots\left(\frac{1}{1}\right) = \frac{1}{n!} \tag{5}$$

Ahora en la suma $\sum P(A_j \cap A_k)$, hay $\binom{n}{2} = {}_nC_2$ términos que tienen el valor dado por (*3*). De igual manera en $\sum P(A_i \cap A_j \cap A_k)$, hay $\binom{n}{3} = {}_nC_3$ términos que tienen el valor dado por (*4*). Por consiguiente, la probabilidad requerida es

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \binom{n}{1}\left(\frac{1}{n}\right) - \binom{n}{2}\left(\frac{1}{n}\right)\left(\frac{1}{n-1}\right) + \binom{n}{3}\left(\frac{1}{n}\right)\left(\frac{1}{n-1}\right)\left(\frac{1}{n-2}\right) - \cdots + (-1)^{n-1}\binom{n}{n}\left(\frac{1}{n!}\right)$$

$$= 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n-1}\,\frac{1}{n!}$$

Del cálculo sabemos que (*véase* apéndice A)

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

así que para $x = -1$

$$e^{-1} = 1 - \left(1 - \frac{1}{2!} + \frac{1}{3!} - \cdots\right)$$

o

$$1 - \frac{1}{2!} + \frac{1}{3!} - \cdots = 1 - e^{-1}$$

Se deduce que si n es grande, la probabilidad requerida es muy cercana a $1 - e^{-1} = 0.6321$. Esto significa que hay una buena posibilidad de que al menos 1 carta llegue al destinatario correcto. El resultado es asombroso debido a que la probabilidad se mantiene prácticamente constante para todos los $n > 10$. Por consiguiente, la probabilidad de que al menos 1 carta llegue al destinatario correcto es prácticamente la misma tanto si n es 10 como si es 10,000.

1.45. Encuentre la probabilidad de que n personas ($n \leqq 365$) seleccionadas al azar tengan n fechas diferentes de cumpleaños.

Suponemos que hay sólo 365 días en un año y que todos los días son igualmente probables, suposiciones que por lo general no se cumplen en la realidad.

La primera de las n personas tiene, por supuesto, una fecha de cumpleaños cuya probabilidad es $365/365 = 1$. Entonces, si la segunda tiene una fecha de cumpleaños diferente, ésta debe ocurrir en alguno de los 364 días restantes. Por lo tanto, la probabilidad de que la segunda persona tenga una fecha de cumpleaños diferente es $364/365$. Igualmente, la probabilidad de que la tercera persona tenga una fecha de cumpleaños diferente de las dos primeras es $363/365$. Por último, la probabilidad de que la n-ésima persona tenga una fecha de cumpleaños diferente de la de las demás personas es $(365 - n + 1)/365$. Por consiguiente tenemos

$$P(\text{todas las } n \text{ fechas de cumpleaños son diferentes}) = \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdots \frac{365 - n + 1}{365}$$

$$= \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right) \cdots \left(1 - \frac{n-1}{365}\right)$$

1.46. Determine cuántas personas se requieren en el problema 1.45 para que la probabilidad de fechas distintas de cumpleaños sea menor que $1/2$.

Denotando la probabilidad dada por p y tomando logaritmos naturales, encontramos

$$(1) \qquad \ln p = \ln\left(1 - \frac{1}{365}\right) + \ln\left(1 - \frac{2}{365}\right) + \cdots + \ln\left(1 - \frac{n-1}{365}\right)$$

Pero sabemos del cálculo (apéndice A, fórmula 7) que

$$(2) \qquad \ln(1 - x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \cdots$$

de manera que (*1*) puede escribirse como

$$(3) \qquad \ln p = -\left[\frac{1 + 2 + \cdots + (n-1)}{365}\right] - \frac{1}{2}\left[\frac{1^2 + 2^2 + \cdots + (n-1)^2}{(365)^2}\right] - \cdots$$

Usando los hechos de que para $n = 2, 3, \ldots$ (apéndice A, fórmulas 1 y 2)

$$(4) \qquad 1 + 2 + \cdots + (n-1) = \frac{n(n-1)}{2}, \qquad 1^2 + 2^2 + \cdots + (n-1)^2 = \frac{n(n-1)(2n-1)}{6}$$

obtenemos para (*3*)

$$\ln p = -\frac{n(n-1)}{730} - \frac{n(n-1)(2n-1)}{12(365)^2} - \cdots \tag{5}$$

Para n pequeño comparado con 365, por ejemplo, $n < 30$, el segundo término y los superiores a la derecha de (5) son despreciables comparados con el primer término, de manera que en este caso una buena aproximación es

$$\ln p = -\frac{n(n-1)}{730} \tag{6}$$

Para $p = \frac{1}{2}$, $\ln p = -\ln 2 = -0.693$. Por consiguiente tenemos

$$\frac{n(n-1)}{730} = 0.693 \quad \text{o} \quad n^2 - n - 506 = 0 \quad \text{o} \quad (n-23)(n+22) = 0 \tag{7}$$

de manera que $n = 23$. Por lo tanto, nuestra conclusión es que si n es mayor que 23, podemos decir con mayor certeza que al menos 2 personas tendrán la misma fecha de cumpleaños.

Problemas suplementarios

CÁLCULO DE PROBABILIDADES

1.47. Determine la probabilidad p, o un estimado de ella, para cada uno de los siguientes eventos:

(*a*) Obtener rey, as, jota de tréboles o reina de diamantes de un naipe bien barajado al sacar sólo una carta.

(*b*) Obtener 8 en la suma del lanzamiento de dos dados balanceados.

(*c*) Encontrar a continuación un tornillo no defectuoso si de 600 que se han examinado se encontraron 12 defectuosos.

(*d*) Obtener 7 u 11 en un lanzamiento de dos dados balanceados.

(*e*) Obtener al menos 1 cara en 3 lanzamientos de una moneda balanceada.

1.48. Un experimento consiste en sacar 3 cartas seguidas de un naipe bien barajado. Sea A_1 el evento "rey en el primer retiro", A_2 "rey en el segundo retiro" y A_3 "rey en el tercer retiro". Diga en palabras el significado de cada una de las siguientes expresiones:

(*a*) $P(A_1 \cap A_2')$, (*b*) $P(A_1 \cup A_2)$, (*c*) $P(A_1' \cup A_2')$, (*d*) $P(A_1' \cap A_2' \cap A_3')$, (*e*) $P[(A_1 \cap A_2) \cup (A_2' \cap A_3)]$.

1.49. Se saca al azar una canica de una caja que contiene 10 canicas rojas, 30 blancas, 20 azules y 14 anaranjadas. Encuentre la probabilidad de que la canica sea (*a*) anaranjada o roja, (*b*) no roja o azul, (*c*) no azul, (*d*) blanca, (*e*) roja, blanca, o azul.

1.50. Se sacan, con remplazo, una tras otra dos canicas de la caja del problema 1.49. Encuentre la probabilidad de que (*a*) ambas sean blancas, (*b*) la primera sea roja y la segunda blanca, (*c*) ninguna sea anaranjada, (*d*) sean roja o blanca o ambas (roja y blanca), (*e*) la segunda no sea azul, (*f*) la primera sea anaranjada, (*g*) al menos una sea azul, (*h*) máximo una sea roja, (*i*) la primera sea blanca, pero la segunda no, (*j*) sólo una sea roja.

1.51. Desarrolle el problema 1.50 sin remplazo después de cada retiro.

PROBABILIDAD CONDICIONAL Y EVENTOS INDEPENDIENTES

1.52. Una caja contiene 2 canicas rojas y 3 azules. Encuentre la probabilidad de que si se sacan dos canicas al azar (sin remplazo) (*a*) ambas sean azules, (*b*) ambas sean rojas, (*c*) una sea roja y la otra azul.

1.53. Encuentre la probabilidad de sacar 3 ases al azar de un naipe de 52 cartas, si las cartas (*a*) se remplazan, (*b*) no se remplazan.

1.54. Si al menos uno de los dos hijos de una familia es varón, ¿cuál es la probabilidad de que ambos hijos sean varones?

1.55. La caja I contiene 3 bolas rojas y 5 blancas, mientras que la caja II contiene 4 bolas rojas y 2 blancas. Se escoge al azar una bola de la primera caja y se coloca en la segunda caja sin haber visto su color. Luego, se retira una bola de la segunda caja. Encuentre la probabilidad de que ésta sea blanca.

TEOREMA O REGLA DE BAYES

1.56. Una caja contiene 3 canicas azules y 2 rojas, mientras que otra contiene 2 azules y 5 rojas. Se retira una canica al azar de una de las dos cajas, y resulta ser azul. ¿Cuál es la probabilidad de que venga de la primera caja?

1.57. Cada uno de tres joyeros idénticos tiene dos cajones. En cada cajón del primer joyero hay un reloj de oro. En cada cajón del segundo joyero hay un reloj de plata. En un cajón del tercer joyero hay un reloj de oro, mientras que en el otro cajón hay un reloj de plata. Si escogemos un joyero al azar, abrimos uno de los cajones y encontramos un reloj de plata, ¿cuál es la probabilidad de que en otro cajón haya un reloj de oro?

1.58. La urna *I* tiene 2 bolas blancas y 3 negras; la urna *II* tiene 4 blancas y 1 negra; y la urna *III* tiene 3 blancas y 4 negras. Se selecciona una urna al azar y se saca una bola, que resulta ser blanca. Encuentre la probabilidad de que se haya seleccionado la urna *I*.

ANÁLISIS COMBINATORIO, CONTEO Y DIAGRAMAS DE ÁRBOL

1.59. Se lanza 3 veces una moneda. Use un diagrama de árbol para determinar las diferentes posibilidades que surgen.

1.60. Se sacan al azar (sin remplazo) tres cartas de una baraja común de 52 cartas. Encuentre el número de maneras como se pueden sacar (*a*) un diamante, un trébol y un corazón, (*b*) dos corazones y luego un trébol o una pica.

1.61. ¿De cuántas maneras se pueden colocar 3 monedas distintas en 2 monederos diferentes?

PERMUTACIONES

1.62. Calcule (*a*) ${}_4P_2$, (*b*) ${}_4P_5$, (*c*) ${}_{10}P_3$

1.63. ¿Para cuál valor de n es ${}_{n+1}P_3 = {}_nP_4$?

1.64. ¿De cuántas maneras posibles se pueden sentar 5 personas en un sofá si solamente hay 3 puestos disponibles?

1.65. ¿De cuántas maneras se pueden ordenar 7 libros en una biblioteca si (*a*) cualquier orden es posible, (*b*) 3 libros en particular deben quedar siempre juntos, (*c*) dos libros en particular deben quedar en las esquinas?

1.66. ¿Cuántos números de cinco dígitos diferentes cada uno se pueden formar con los números 1, 2, 3, ..., 9 si (*a*) los números deben ser impares, (*b*) los dos primeros dígitos de cada número son pares?

1.67. Resuelva el problema 1.66 si se permiten repeticiones de los números.

1.68. ¿Cuántos números diferentes de tres dígitos se pueden formar con 3 cuatros, 4 dos, y 2 tres?

1.69. ¿De cuántas maneras pueden sentarse 3 hombres y 3 mujeres alrededor de una mesa redonda si (*a*) no hay restricciones, (*b*) 2 mujeres en particular no pueden sentarse juntas, (*c*) cada mujer se encuentra entre 2 hombres?

COMBINACIONES

1.70. Calcule (*a*) ${}_5C_3$, (*b*) ${}_8C_4$, (*c*) ${}_{10}C_8$.

1.71. ¿Para qué valor de n es $3 \cdot {}_{n+1}C_3 = 7 \cdot {}_nC_2$?

1.72. ¿De cuántas maneras se pueden escoger 6 preguntas de 10?

1.73. ¿Cuántos comités diferentes de 3 hombres y 4 mujeres se pueden formar a partir de 8 hombres y 6 mujeres?

1.74. ¿De cuántas maneras se pueden seleccionar 2 hombres, 4 mujeres, 3 niños y 3 niñas a partir de 6 hombres, 8 mujeres, 4 niños y 5 niñas si (*a*) no se imponen restricciones, (*b*) se deben escoger un hombre y una mujer en particular?

1.75. ¿De cuántas maneras puede dividirse un grupo de 10 personas en (*a*) dos grupos de 7 y 3 personas, (*b*) tres grupos de 5, 3, y 2 personas?

1.76. A partir de un grupo de 5 estadísticos y 6 economistas se va a formar un comité de 3 estadísticos y 2 economistas. ¿Cuántos comités diferentes se pueden formar si (*a*) no se imponen restricciones, (*b*) 2 estadísticos en particular deben estar en el comité, (*c*) 1 economista en particular no puede estar en el comité?

1.77. Encuentre el número de (*a*) combinaciones y (*b*) permutaciones de 4 letras cada una que pueden formarse con las letras de la palabra *Tennessee*.

COEFICIENTES BINOMIALES

1.78. Calcule (*a*) ${}_6C_3$, (*b*) $\binom{11}{4}$, (*c*) $({}_8C_2)({}_4C_3)/{}_{12}C_5$.

1.79. Expanda (*a*) $(x+y)^6$, (*b*) $(x-y)^4$, (*c*) $(x-x^{-1})^5$, (*d*) $(x^2+2)^4$.

1.80. Encuentre el coeficiente de x en $\left(x+\frac{2}{x}\right)^9$.

PROBABILIDAD USANDO EL ANÁLISIS COMBINATORIO

1.81. En dos lanzamientos de un par de dados balanceados encuentre la probabilidad de obtener un puntaje de 7 (*a*) una vez, (*b*) al menos una vez, (*c*) dos veces.

1.82. Se sacan una tras otra dos cartas de un naipe común y bien barajado de 52 cartas. Encuentre la probabilidad de que (*a*) la primera carta no sea un diez de tréboles o un as, (*b*) la primera carta sea un as, pero la segunda no, (*c*) al menos una carta sea de diamantes, (*d*) las cartas no sean del mismo palo, (*e*) no más de una carta sea figura (jota, reina, rey), (*f*) la segunda carta no sea figura, (*g*) la segunda carta no sea figura, dado que la primera sí lo fue, (*h*) las cartas sean figuras o picas, o ambas.

1.83. Una caja contiene 9 boletas numeradas del 1 al 9 inclusive. Si se sacan una a una tres boletas de la caja, encuentre la probabilidad de que de manera alternada sean impares, pares, impares o pares, impares, pares.

1.84. Las apuestas a favor de que *A* le gane a *B* un juego de ajedrez están 3:2. Si se van a jugar 3 partidas, ¿cuáles serán las apuestas (*a*) a favor de que *A* gane al menos 2 de 3 juegos, (*b*) en contra de *A* perdiendo los primeros dos juegos contra *B*?

1.85. En un juego de Bridge, cada uno de los 4 jugadores recibe 13 cartas de un naipe bien barajado de 52 cartas. Encuentre la probabilidad de que un jugador (digamos, el mayor) reciba (*a*) 7 de diamantes, 2 de tréboles, 3 de corazones y 1 de picas; (*b*) un palo completo.

1.86. De una urna que contiene 6 canicas rojas y 8 azules se retiran al azar cinco canicas sin remplazo. Encuentre la probabilidad de que 3 sean rojas y 2 azules.

1.87. (*a*) Encuentre la probabilidad de obtener la suma de 7 en al menos 1 de 3 lanzamientos de un par de dados balanceados. (*b*) ¿Cuántos lanzamientos se necesitan para que la probabilidad en (*a*) sea mayor de 0.95?

1.88. Se sacan tres cartas de un naipe común de 52 cartas. Encuentre la probabilidad de que (*a*) todas las cartas sean de un palo, (*b*) al menos se saquen 2 ases.

1.89. Encuentre la probabilidad de que un jugador de Bridge reciba 13 cartas de las cuales 9 pertenezcan a un solo palo.

PROBLEMAS DIVERSOS

1.90. Un espacio muestral consta de 3 puntos muestrales con probabilidades asociadas dadas por $2p$, p^2 y $4p - 1$. Encuentre el valor de p.

1.91. ¿Cuántas palabras pueden formarse con 5 letras si (*a*) todas las letras son diferentes, (*b*) 2 letras son idénticas, (*c*) todas las letras son distintas, pero 2 en particular no pueden ser adyacentes?

1.92. Se escogen al azar cuatro números enteros entre 0 y 9, inclusive. Encuentre la probabilidad de que (*a*) todos sean diferentes, (*b*) no más de 2 sean iguales.

1.93. Se lanza un par de dados repetidamente. Encuentre la probabilidad de que por primera vez ocurra un 11 en el sexto lanzamiento.

..94. ¿Cuál es el menor número de lanzamientos necesarios en el problema 1.93 de manera que la probabilidad de obtener 11 sea mayor de (*a*) 0.5, (*b*) 0.95?

1.95. Encuentre la probabilidad de obtener en un juego de póker (*a*) una escalera flor, la cual conste de diez, jota, reina, rey y as del mismo palo, (*b*) un fuljan, que conste de 3 cartas de un valor y 2 de otro (como 3 diez y 2 jotas), (*c*) todas las cartas diferentes, (*d*) 4 ases.

1.96. La probabilidad de que un hombre acierte al blanco es 2/3. Si él le dispara al blanco hasta que acierta por primera vez, encuentre la probabilidad de que le tomen 5 disparos para acertar al blanco.

1.97. (*a*) Una estantería contiene 6 compartimentos separados. ¿De cuántas maneras se pueden colocar 4 canicas idénticas en los compartimentos? (*b*) Desarrolle el problema si hay n compartimentos y r canicas. Este tipo de problemas surge en física en relación con las *estadísticas de Bose-Einstein*.

1.98. (*a*) Una estantería contiene 6 compartimentos separados. ¿De cuántas maneras se pueden colocar 12 canicas idénticas, de manera que no haya compartimentos vacíos? (*b*) Desarrolle el problema si hay n compartimentos y r canicas donde $r > n$. Este tipo de problemas surge en física en relación con las *estadísticas de Fermi-Dirac*.

1.99. Un jugador de póker tiene las cartas 2, 3, 4, 6, 8. Él quiere desencartarse del 8 y remplazarlo por otra carta, la cual él espera que sea un 5 (caso en el cual tendría una escalera). ¿Cuál es la probabilidad de que la obtenga, suponiendo que los otros tres jugadores tengan juntos (*a*) un cinco, (*b*) dos cincos, (*c*) tres cincos, (*d*) ningún cinco? ¿Se puede desarrollar el problema si se desconoce el número de cincos que tienen los otros jugadores en sus manos? Explique.

1.100. Desarrolle el problema 1.40 si el juego se limita a 3 lanzamientos.

1.101. Encuentre la probabilidad de que en un juego de Bridge (*a*) 2, (*b*) 3, (*c*) los 4 jugadores tengan un palo completo.

RESPUESTAS A LOS PROBLEMAS SUPLEMENTARIOS

1.47. (*a*) 5/26 (*b*) 5/36 (*c*) 0.98 (*d*) 2/9 (*e*) 7/8

1.48. (*a*) Probabilidad de rey en el primer retiro y no rey en el segundo.

(*b*) Probabilidad de rey en el primer retiro, o rey en el segundo retiro o ambos.

(*c*) No rey en el primer retiro, o no rey en el segundo retiro o ambos (no rey en el primero y segundo retiros).

(*d*) No rey en el primero, segundo y tercer retiros.

(*e*) Probabilidad de rey en el primer retiro y rey en el segundo retiro o no rey en el segundo retiro y rey en el tercer retiro.

1.49. (*a*) 1/3 (*b*) 3/5 (*c*) 11/15 (*d*) 2/5 (*e*) 4/5

1.50. (*a*) 4/25 (*c*) 16/25 (*e*) 11/15 (*g*) 104/225 (*i*) 6/25
(*b*) 4/75 (*d*) 64/225 (*f*) 1/5 (*h*) 221/225 (*j*) 52/225

1.51. (*a*) 29/185 (*c*) 118/185 (*e*) 11/15 (*g*) 86/185 (*i*) 9/37
(*b*) 2/37 (*d*) 52/185 (*f*) 1/5 (*h*) 182/185 (*j*) 26/111

1.52. (*a*) 3/10 (*b*) 1/10 (*c*) 3/5

1.53. (*a*) 1/2197 (*b*) 1/17,576

1.54. 1/3

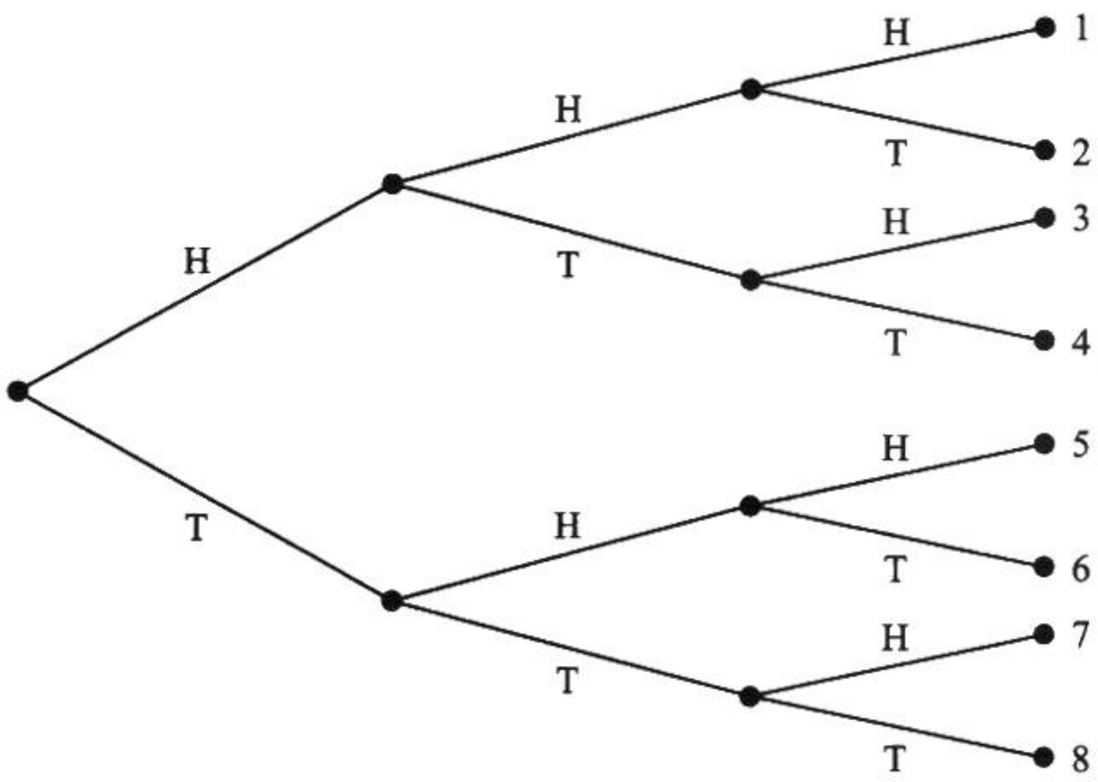
1.55. 21/56

1.56. 21/31

1.57. 1/3

1.58. 14/57

1.59.



1.60. (*a*) $13 \times 13 \times 13$ (*b*) $13 \times 12 \times 26$

1.61. 8

1.62. (*a*) 12 (*b*) 2520 (*c*) 720

1.63. $n = 5$

1.64. 60

1.65. (*a*) 5040 (*b*) 720 (*c*) 240

1.66. (*a*) 8400 (*b*) 2520

1.67. (*a*) 32,805 (*b*) 11,664

1.68. 26

1.69. (*a*) 120 (*b*) 72 (*c*) 12

1.70. (*a*) 10 (*b*) 70 (*c*) 45

1.71. $n = 6$

1.72. 210

1.73. 840

1.74. (*a*) 42,000 (*b*) 7000

1.75. (*a*) 120 (*b*) 2520

1.76. (*a*) 150 (*b*) 45 (*c*) 100

1.77. (*a*) 17 (*b*) 163

1.78. (*a*) 20 (*b*) 330 (*c*) 14/99

1.79. (*a*) $x^6 + 6x^5y + 15x^4y^2 + 20x^3y^3 + 15x^2y^3 + 6xy^5 + y^6$
(*b*) $x^4 - 4x^3y + 6x^2y^2 - 4xy^3 + y^4$
(*c*) $x^5 - 5x^3 + 10x - 10x^{-1} + 5x^{-3} - x^{-5}$
(*d*) $x^8 + 8x^6 + 24x^4 + 32x^2 + 16$

1.80. 2016

1.81. (*a*) 5/18 (*b*) 11/36 (*c*) 1/36

1.82. (*a*) 47/52 (*b*) 16/221 (*c*) 15/34 (*d*) 13/17 (*e*) 210/221 (*f*) 10/13 (*g*) 40/51 (*h*) 77/442

1.83. 5/18

1.84. (*a*) 81 : 44 (*b*) 21 : 4

1.85. (*a*) $({}_{13}C_7)({}_{13}C_2)({}_{13}C_3)({}_{13}C_1)/{}_{52}C_{13}$ (*b*) $4/{}_{52}C_{13}$

1.86. $({}_6C_3)({}_8C_2)/{}_{14}C_5$

1.87. (*a*) 91/216 (*b*) al menos 17

1.88. (*a*) $4 \cdot {}_{13}C_3/{}_{52}C_3$ (*b*) $({}_4C_2 \cdot {}_{48}C_1 + {}_4C_3)/{}_{52}C_3$

1.89. $4({}_{13}C_9)({}_{39}C_4)/{}_{52}C_{13}$

1.90. $\sqrt{11} - 3$

1.91. (*a*) 120 (*b*) 60 (*c*) 72

1.92. (*a*) 63/125 (*b*) 963/1000

1.93. 1,419,857/34,012,224

1.94. (*a*) 13 (*b*) 53

1.95. (*a*) $4/{}_{52}C_5$ (*c*) $4^5({}_{13}C_5)/{}_{52}C_5$
(*b*) $(13)(2)(4)(6)/{}_{52}C_5$ (*d*) (5)(4)(3)(2)/(52)(51)(50)(49)

1.96. 2/243

1.97. (*a*) 126 (*b*) ${}_{n+r-1}C_{n-1}$

1.98. (*a*) 462 (*b*) ${}_{r-1}C_{n-1}$

1.99. (*a*) 3/32 (*b*) 1/16 (*c*) 1/32 (*d*) 1/8

1.100. prob. A gane = 61/216, prob. B gane = 5/36, prob. de empate = 125/216

1.101. (*a*) $12/({}_{52}C_{13})({}_{39}C_{13})$ (*b*) $24/({}_{52}C_{13})({}_{39}C_{13})({}_{26}C_{13})$

CAPÍTULO 2

Variables aleatorias y distribución de probabilidad

VARIABLES ALEATORIAS

Supongamos que a cada punto del espacio muestral le asignamos un número. Tenemos entonces una *función* definida en el espacio muestral. Esta función recibe el nombre de *variable aleatoria* (o *variable estocástica*) o más precisamente *función aleatoria* (*función estocástica*). Usualmente se denota con una letra mayúscula, como X o Y. En general, una variable aleatoria tiene alguna significación física, geométrica o de otro tipo.

EJEMPLO 2.1. Supongamos que se lanza una moneda dos veces, de manera que el espacio muestral es $S = \{CC, CS, SC, SS\}$. Siendo X el número de caras que pueden salir. Podemos asociar cada punto muestral con un número de X, como se muestra en la tabla 2-1. Así, por ejemplo, en el caso de CC (es decir, 2 caras), $X = 2$, mientras que para CS (1 cara), $X = 1$. Concluimos entonces que X es una variable aleatoria.

Tabla 2-1

Punto muestral	CC	CS	SC	SS
X	2	1	1	0

Se debe observar que muchas otras variables aleatorias también pueden definirse en este espacio muestral, por ejemplo, el cuadrado del número de caras o el número de caras menos el número de sellos.

Una variable aleatoria que toma un número finito o contable infinito de valores (*véase* página 4) se llama una *variable aleatoria discreta*, mientras que una que toma un número de valores infinito no contable recibe el nombre de *variable aleatoria no discreta*.

DISTRIBUCIONES DE PROBABILIDAD DISCRETA

Sea X una variable aleatoria discreta, y supongamos que los valores posibles que ésta puede asumir están dados por $x_1, x_2, x_3, \ldots$, con determinado orden. Supongamos también que estos valores se asumen con probabilidades dadas por

$$P(X = x_k) = f(x_k) \qquad k = 1, 2, \ldots. \tag{1}$$

Es conveniente presentar la *función de probabilidad*, también denominada *distribución de probabilidad*, dada por

$$P(X = x) = f(x) \tag{2}$$

Para $x = x_k$, se reduce a (*1*), mientras que para otros valores de x, $f(x) = 0$.

En general, $f(x)$ es una función de probabilidad si

1. $f(x) \geqq 0$
2. $\sum_x f(x) = 1$

donde la suma en 2 toma todos los valores posibles de x.

EJEMPLO 2.2. Encuentre la función de probabilidad correspondiente a la variable aleatoria X del ejemplo 2.1.

Suponiendo que la moneda es balanceada, tenemos

$$P(CC) = \frac{1}{4} \qquad P(CS) = \frac{1}{4} \qquad P(SC) = \frac{1}{4} \qquad P(SS) = \frac{1}{4}$$

Entonces

$$P(X = 0) = P(SS) = \frac{1}{4}$$

$$P(X = 1) = P(CS \cup SC) = P(CS) + P(SC) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

$$P(X = 2) = P(CC) = \frac{1}{4}$$

La función de probabilidad está dada, entonces, por la tabla 2-2.

Tabla 2-2

x	0	1	2
$f(x)$	1/4	1/2	1/4

FUNCIONES DE DISTRIBUCIÓN PARA VARIABLES ALEATORIAS

La *función de distribución acumulada*, o de manera breve la *función de distribución*, para una variable aleatoria X está definida por

$$F(x) = P(X \leq x) \tag{3}$$

donde x es un número real cualquiera, es decir, $-\infty < x < \infty$.

La función de distribución $F(x)$ tiene las siguientes propiedades:

1. $F(x)$ no es decreciente [es decir, $F(x) \leq F(y)$ si $x \leq y$].
2. $\lim_{x \to -\infty} F(x) = 0$; $\lim_{x \to \infty} F(x) = 1$.

3. $F(x)$ es continua desde la derecha [o sea, $\lim_{h\to 0^+} F(x+h) = F(x)$ para todo x].

FUNCIONES DE DISTRIBUCIÓN PARA VARIABLES ALEATORIAS DISCRETAS

La función de distribución para una variable aleatoria discreta X puede obtenerse a partir de su función de probabilidad notando que, para todo x en $(-\infty, \infty)$,

$$F(x) = P(X \leq x) = \sum_{u \leq x} f(u) \tag{4}$$

donde la suma sustituye todo los valores u tomados por X para la cual $u \leq x$.

Si X toma solamente un número finito de valores $x_1, x_2, \ldots, x_n$, entonces la función de distribución está dada por

$$F(x) = \begin{cases} 0 & -\infty < x < x_1 \\ f(x_1) & x_1 \leqq x < x_2 \\ f(x_1) + f(x_2) & x_2 \leqq x < x_3 \\ \vdots & \vdots \\ f(x_1) + \cdots + f(x_n) & x_n \leqq x < \infty \end{cases} \tag{5}$$

EJEMPLO 2.3. (a) Encuentre la función de distribución para la variable aleatoria X del ejemplo 2.2. (b) Elabore su gráfica.

(a) La función de distribución es

$$F(x) = \begin{cases} 0 & -\infty < x < 0 \\ \frac{1}{4} & 0 \leqq x < 1 \\ \frac{3}{4} & 1 \leqq x < 2 \\ 1 & 2 \leqq x < \infty \end{cases}$$

(b) La gráfica para $F(x)$ se muestra en la figura 2-1.

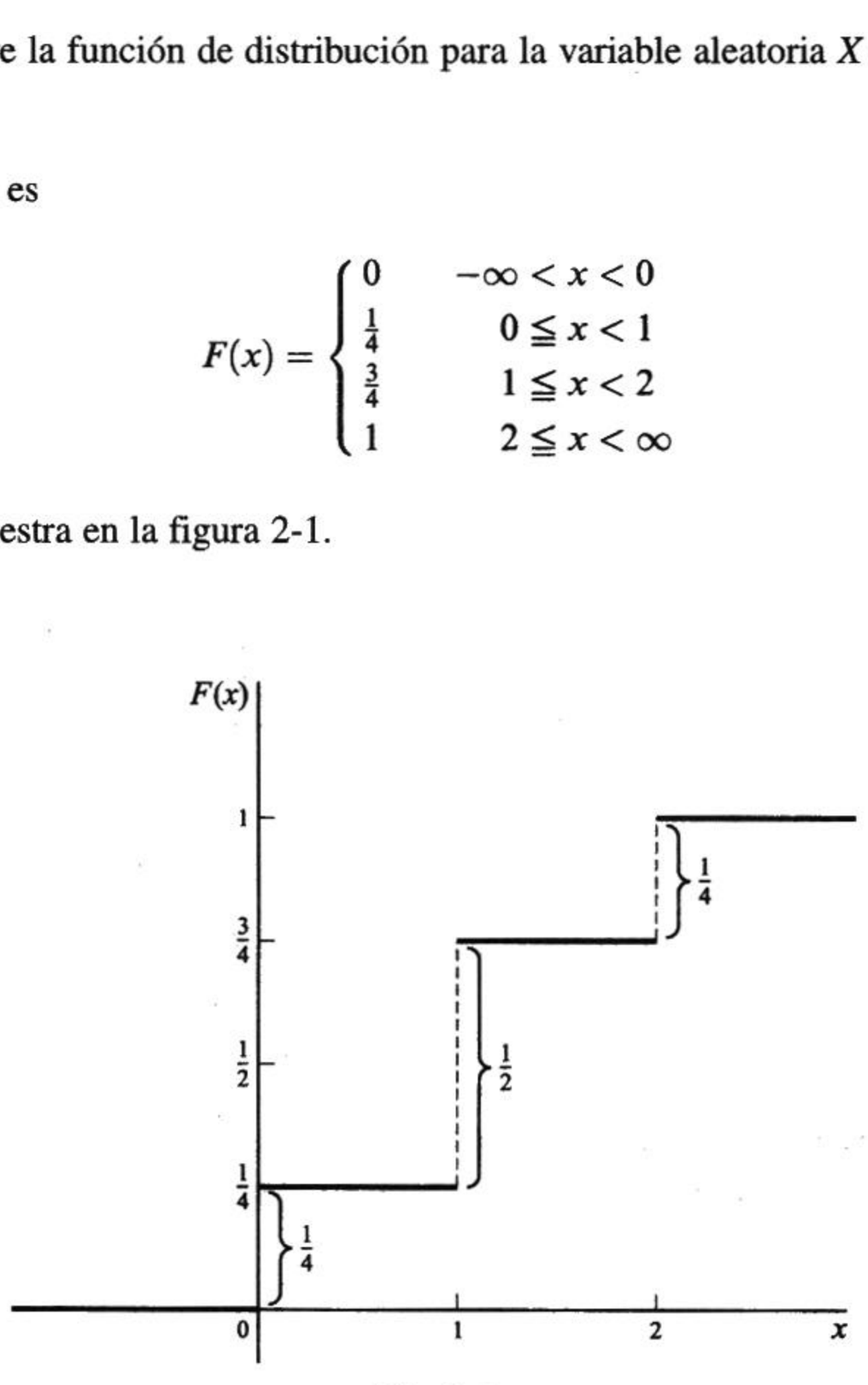


Fig. 2-1

Observe los siguientes aspectos acerca de la distribución anterior, los cuales son verdaderos.

1. Las magnitudes de los saltos en 0, 1, 2, son $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$ que son precisamente las probabilidades en la tabla 2-2. Este hecho permite obtener la función de probabilidad a partir de la función de distribución.
2. Debido a la apariencia de la gráfica de la figura. 2-1, ésta se denomina con frecuencia una *función escalonada* o *función paso*. El valor de la función en un entero se obtiene a partir del paso más grande; así, el valor en 1 es $\frac{3}{4}$ y no $\frac{1}{4}$. Esto se expresa matemáticamente estableciendo que la función de distribución es *continua a la derecha* en 0, 1, 2.
3. A medida que avanzamos de izquierda a derecha (o sea, *subiendo la escalera*), la función de distribución permanece igual o aumenta, tomando valores entre 0 y 1. Debido a esto, se dice que es una *función monótonamente creciente*.

A partir de la observación anterior y de las propiedades de las funciones de distribución, es claro que la función de probabilidad de una variable aleatoria discreta puede obtenerse a partir de la función de distribución, notando que

$$f(x) = F(x) - \lim_{u \to x^-} F(u). \tag{6}$$

VARIABLES ALEATORIAS CONTINUAS

Se dice que una variable aleatoria no discreta X es *absolutamente continua*, o simplemente *continua*, si su función de distribución se puede representar como

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\, du \qquad (-\infty < x < \infty) \tag{7}$$

donde la función $f(x)$ tiene las propiedades

1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$

A partir de lo anterior se deduce que si X es una variable aleatoria continua, entonces la probabilidad de que X tome cualquier valor particular es cero, mientras que la *probabilidad de intervalo* de que X se encuentre entre dos valores diferentes, por ejemplo a y b, está dada por

$$P(a < X < b) = \int_{a}^{b} f(x)\, dx \tag{8}$$

EJEMPLO 2.4. Si un individuo se selecciona al azar de un grupo grande de hombres adultos, la probabilidad de que su altura X sea precisamente 180 centímetros (es decir, 180.000... centímetros) es cero. Sin embargo, hay una probabilidad mayor que cero de que X se encuentre entre 175.000... centímetros y 180.500 centímetros, por ejemplo.

Una función $f(x)$ que satisfaga las necesidades anteriores se llama *función de probabilidad* o *distribución de probabilidad* para una variable aleatoria continua, pero con mayor frecuencia se denomina *función de densidad de probabilidad* o simplemente *función de densidad*. Cualquier función $f(x)$ que satisfaga las propiedades anteriores 1 y 2 será automáticamente una función de densidad, y las probabilidades requeridas se obtendrán a partir de (*8*).

EJEMPLO 2.5. (*a*) Encuentre la constante c de manera que la función

$$f(x) = \begin{cases} cx^2 & 0 < x < 3 \\ 0 & \text{de otra manera} \end{cases}$$

es una función de densidad, y (*b*) calcule $P(1 < X < 2)$.

(*a*) Dado que $f(x)$ satisface la Propiedad 1 si $c \geqq 0$, debe satisfacer la Propiedad 2 con el fin de que sea una función de densidad. Ahora

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_0^3 cx^2\,dx = \frac{cx^3}{3}\Big|_0^3 = 9c$$

y dado que esto debe ser igual a 1, tenemos que $c = 1/9$.

(*b*)
$$P(1 < X < 2) = \int_1^2 \frac{1}{9}x^2\,dx = \frac{x^3}{27}\Big|_1^2 = \frac{8}{27} - \frac{1}{27} = \frac{7}{27}$$

En caso de que $f(x)$ sea una variable continua, lo cual debemos suponer a menos que se especifique otra cosa, la probabilidad de que X es igual a cualquier valor particular es cero. En tal caso, podemos remplazar uno o ambos de los signos $<$ en (*8*) por $\leqq$. Así, en el ejemplo 2.5,

$$P(1 \leqq X \leqq 2) = P(1 \leqq X < 2) = P(1 < X \leqq 2) = P(1 < X < 2) = \frac{7}{27}$$

EJEMPLO 2.6. (*a*) Encuentre la función de distribución para la variable aleatoria del ejemplo 2.5. (*b*) Use el resultado de (*a*) para encontrar $P(1 < x \leqq 2)$.

(*a*) Tenemos

$$F(x) = P(X \leqq x) = \int_{-\infty}^{x} f(u)\,du$$

Si $x < 0$, entonces $F(x) = 0$. Si $0 \leqq x < 3$, entonces

$$F(x) = \int_0^x f(u)\,du = \int_0^x \frac{1}{9}u^2\,du = \frac{x^3}{27}$$

Si $x \geqq 3$, entonces

$$F(x) = \int_0^3 f(u)\,du + \int_3^x f(u)\,du = \int_0^3 \frac{1}{9}u^2\,du + \int_3^x 0\,du = 1$$

Así la función de distribución es

$$F(x) = \begin{cases} 0 & x < 0 \\ x^3/27 & 0 \leqq x < 3 \\ 1 & x \geqq 3 \end{cases}$$

Observe que $F(x)$ aumenta monótonamente desde 0 a 1, tal como se requiere para una función de distribución. También se debe observar que en este caso $F(x)$ es continua.

(*b*) Tenemos

$$\begin{aligned} P(1 < X \leqq 2) &= P(X \leqq 2) - P(X \leqq 1) \\ &= F(2) - F(1) \\ &= \frac{2^3}{27} - \frac{1^3}{27} = \frac{7}{27} \end{aligned}$$

como en el ejemplo 2.5.

La probabilidad de que X esté entre x y $x + \Delta x$ está dada por

$$P(x \leqq X \leqq x + \Delta x) = \int_x^{x+\Delta x} f(u)\, du \tag{9}$$

De manera que si Δx es pequeño, tenemos aproximadamente

$$P(x \leqq X \leqq x + \Delta x) = f(x)\Delta x \tag{10}$$

También vemos a partir de (7) que al diferenciar ambos lados

$$\frac{dF(x)}{dx} = f(x) \tag{11}$$

en todos los puntos donde $f(x)$ es continua; es decir, la derivada de la función de distribución es la función de densidad.

Se debe señalar que existen variables aleatorias que no son discretas ni continuas. Se puede demostrar que es un ejemplo la variable aleatoria X con la siguiente función de distribución.

$$F(x) = \begin{cases} 0 & x < 1 \\ \dfrac{x}{2} & 1 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$$

Con el fin de obtener (*11*), usamos la propiedad básica

$$\frac{d}{dx}\int_a^x f(u)\, du = f(x) \tag{12}$$

la cual es una versión del teorema fundamental del cálculo.

INTERPRETACIONES GRÁFICAS

Si $f(x)$ es la función de densidad para una variable aleatoria X, entonces podemos representar gráficamente $y = f(x)$ por medio de una curva, tal como aparece en la figura 2-2. Dado que $f(x) \geqq 0$, la curva no puede caer por debajo del eje x. El área completa limitada por la curva y el eje x debe ser igual a 1, debido a la Propiedad 2 de la página 41. Geométricamente, la probabilidad de que X se encuentre entre a y b, es decir, $P(a \leq X \leq b)$, se representa por el área sombreada de la figura 2-2.

La función de distribución $F(x) = P(X \leqq x)$ es una función monótonamente creciente, que se incrementa de 0 a 1 y está representada por una curva, tal como aparece en la figura 2-3.

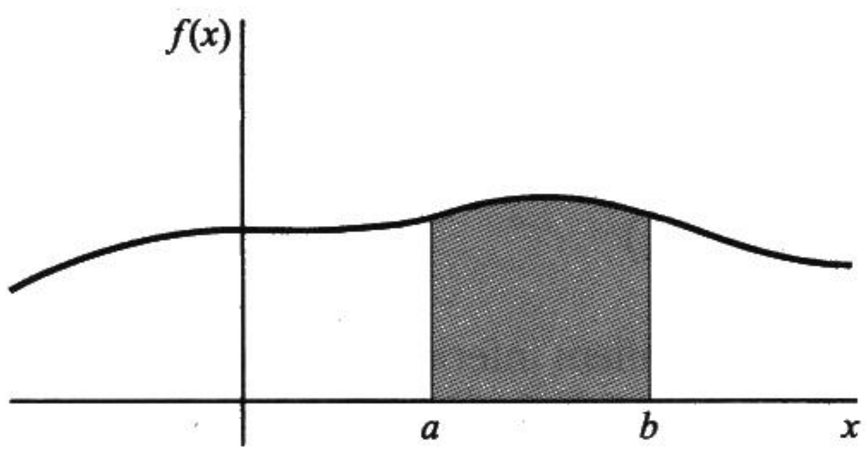


Fig. 2-2

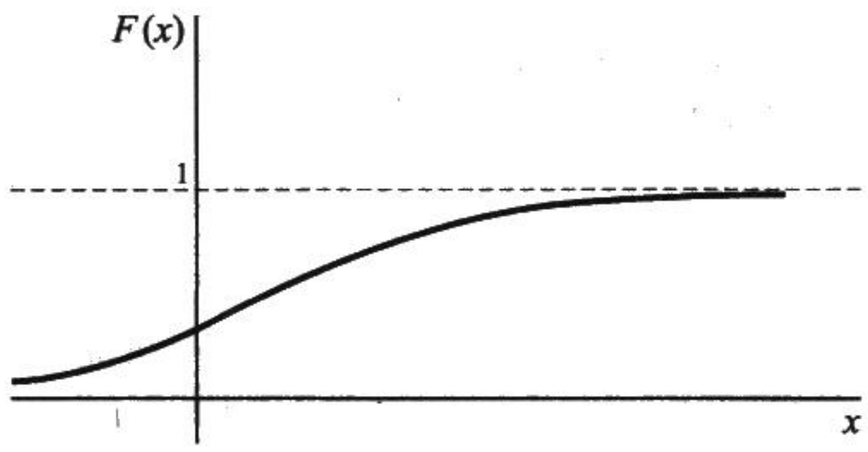


Fig. 2-3

DISTRIBUCIONES CONJUNTAS

Los conceptos anteriores se generalizan fácilmente como dos o más variables aleatorias. Consideramos el caso típico de dos variables aleatorias que son discretas o continuas. En los casos donde una variable es discreta y la otra continua, se pueden hacer modificaciones con facilidad. También es posible hacer generalizaciones para más de dos variables.

1. CASO DISCRETO. Si X y Y son dos variables aleatorias discretas, definimos la función de probabilidad conjunta de X y Y por

$$P(X = x, Y = y) = f(x, y) \tag{13}$$

donde 1. $f(x, y) \geqq 0$

2. $\sum_x \sum_y f(x, y) = 1$

es decir, la suma de todos los valores de x y y es igual a uno.

Supongamos que X puede asumir cualquiera de los m valores $x_1, x_2, \ldots, x_m$ y Y puede asumir cualquiera de los n valores $y_1, y_2, \ldots, y_n$. Entonces la probabilidad del evento $X = x_j$ y $Y = y_k$ está dada por

$$P(X = x_j, Y = y_k) = f(x_j, y_k) \tag{14}$$

La función de probabilidad conjunta para X y Y se puede representar en una tabla de probabilidad conjunta, como la Tabla 2-3. La probabilidad de $X = x_j$ se obtiene sumando todas las entradas en la fila correspondiente a xj, y está dada por

$$P(X = x_j) = f_1(x_j) = \sum_{k=1}^{n} f(x_j, y_k) \tag{15}$$

Tabla 2-3

X \ Y	y_1	y_2	$\cdots$	y_n	Totales ↓
x_1	$f(x_1, y_1)$	$f(x_1, y_2)$	$\cdots$	$f(x_1, y_n)$	$f_1(x_1)$
x_2	$f(x_2, y_1)$	$f(x_2, y_2)$	$\cdots$	$f(x_2, y_n)$	$f_1(x_2)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_m	$f(x_m, y_1)$	$f(x_m, y_2)$	$\cdots$	$f(x_m, y_n)$	$f_1(x_m)$
Totales →	$f_2(y_1)$	$f_2(y_2)$	$\cdots$	$f_2(y_n)$	1 ← Gran total

Para $j = 1, 2, \ldots, m$, éstas se indican por la entrada de totales en la columna de la extrema derecha o margen de la tabla 2-3. De igual manera, la probabilidad de $Y = y_k$ se obtiene sumando todas las entradas en la columna correspondiente a y_k y está dada por

$$P(Y = y_k) = f_2(y_k) = \sum_{j=1}^{m} f(x_j, y_k) \tag{16}$$

Para $k = 1, 2, \ldots, n$, éstas se indican por la entrada de totales en la hilera inferior o margen de la tabla 2-3.

Debido a que las probabilidades (*15*) y (*16*) se obtuvieron a partir de los márgenes de la tabla, con frecuencia nos referimos a $f_1(x_j)$ y $f_2(y_k)$ [o simplemente $f_1(x)$ y $f_2(y)$] como funciones de probabilidad marginal de X y Y, respectivamente. También se debe notar que

$$\sum_{j=1}^{m} f_1(x_j) = 1 \qquad \sum_{k=1}^{n} f_2(y_k) = 1 \tag{17}$$

que puede escribirse

$$\sum_{j=1}^{m} \sum_{k=1}^{n} f(x_j, y_k) = 1 \tag{18}$$

Esto es tan sólo la proposición de que la probabilidad total de todas las entradas es 1. El *gran total* de 1 se indica en la esquina inferior derecha de la tabla.

La *función de distribución conjunta* de X y Y se define como

$$F(x, y) = P(X \leqq x,\ Y \leqq y) = \sum_{u \leqq x} \sum_{v \leqq y} f(u, v) \tag{19}$$

En la tabla 2-3, $F(x, y)$ es la suma de todas las entradas para las cuales $x_j \leqq x$ y $y_k \leqq y$.

2. CASO CONTINUO. El caso en que ambas variables son continuas se obtiene fácilmente por analogía con el caso discreto, remplazando las sumas por integrales. Así, la *función de probabilidad conjunta* para las variables aleatorias X y Y (o como se denominan comúnmente, la *función de densidad conjunta* de X y Y) se define por

1. $f(x, y) \geqq 0$
2. $\displaystyle\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$

Gráficamente $z = f(x, y)$ representa una superficie, llamada *superficie de probabilidad*, como se indica en la figura 2-4. El volumen total limitado por esta superficie y el plano xy es igual a 1 de acuerdo con la Propiedad 2 de arriba. La probabilidad de que X se encuentre entre a y b, mientras que Y esté entre c y d, está dada gráficamente por el volumen sombreado de la figura 2-4, y matemáticamente por

$$P(a < X < b,\ c < Y < d) = \int_{x=a}^{b} \int_{y=c}^{d} f(x, y)\, dx\, dy \tag{20}$$

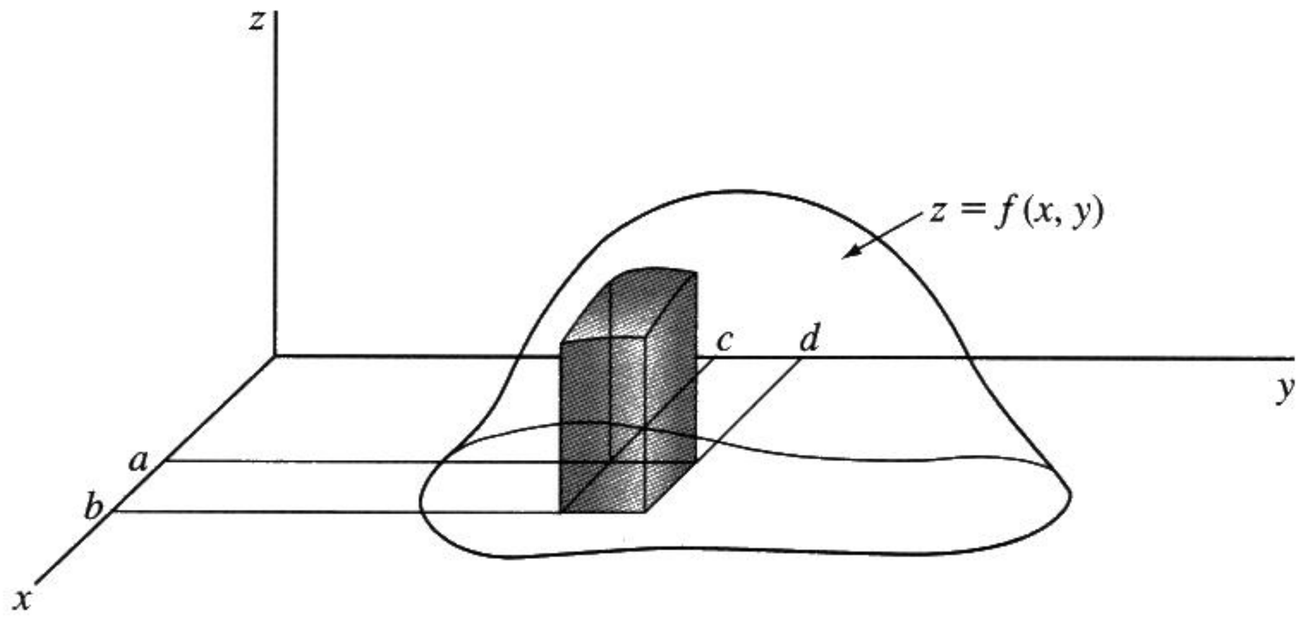


Fig. 2-4

De manera más general, si A representa cualquier evento, habrá una región $\mathcal{R}_A$ del plano xy que le corresponda. En tal caso podemos encontrar la probabilidad de A desarrollando la integral de $\mathcal{R}_A$, es decir.

$$P(A) = \iint_{\mathscr{R}_A} f(x, y)\, dx\, dy \tag{21}$$

La *función de distribución conjunta* de X y Y se define en este caso por

$$F(x, y) = P(X \leqq x,\ Y \leqq y) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{y} f(u, v)\, du\, dv \tag{22}$$

Por analogía con (*11*), página 43, se deduce que

$$\frac{\partial^2 F}{\partial x\, \partial y} = f(x, y) \tag{23}$$

es decir, la función de densidad se obtiene derivando la función de distribución con respecto a x y y.

A partir de (*22*) obtenemos

$$P(X \leqq x) = F_1(x) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{\infty} f(u, v)\, du\, dv \tag{24}$$

$$P(Y \leqq y) = F_2(y) = \int_{u=-\infty}^{\infty} \int_{v=-\infty}^{y} f(u, v)\, du\, dv \tag{25}$$

Llamamos a (*24*) y (*25*) *funciones de distribución marginal*, o simplemente *funciones de distribución* de X y Y respectivamente. Las derivadas de (*24*) y (*25*) con respecto a x y y se denominan entonces las funciones de densidad marginal, o simplemente funciones de densidad de X y Y, y están dadas por

$$f_1(x) = \int_{v=-\infty}^{\infty} f(x, v)\, dv \qquad f_2(y) = \int_{u=-\infty}^{\infty} f(u, y)\, du \tag{26}$$

VARIABLES ALEATORIAS INDEPENDIENTES

Supongamos que X y Y son variables aleatorias discretas. Si los eventos $X = x$ y $Y = y$ son independientes para todo x y y, entonces decimos que X y Y son *variables aleatorias independientes*. En tal caso,

$$P(X = x, Y = y) = P(X = x)\, P(Y = y) \tag{27}$$

o de manera equivalente

$$f(x, y) = f_1(x)\, f_2(y) \tag{28}$$

Inversamente, si para todo x y y la función de probabilidad conjunta $f(x, y)$ puede expresarse sólo como el producto de una función de x sola y como una función de y sola (las cuales son entonces funciones de probabilidad marginal de X y Y), X y Y son independientes. Sin embargo, si $f(x, y)$ no puede expresarse de tal manera, entonces X y Y son *dependientes*.

Si X y Y son variables aleatorias continuas, decimos que son *variables aleatorias independientes* si los eventos $X \leqq x$ y $Y \leqq y$ son eventos independientes para todo x y y. En tal caso podemos escribir

$$P(X \leqq x, Y \leqq y) = P(X \leqq x)\, P(Y \leqq y) \tag{29}$$

o de manera equivalente

$$F(x, y) = F_1(x)\, F_2(y) \tag{30}$$

donde $F_1(z)$ y $F_2(y)$ son las funciones de distribución (marginal) de X y Y respectivamente. Inversamente, X y Y son variables aleatorias independientes si para todo x y y, su función de distribución conjunta $F(x, y)$ puede

expresarse como el producto de la función x sola y de la función y sola (las cuales son las distribuciones marginales de X y Y, respectivamente). Sin embargo, si $F(x, y)$ no puede expresarse de tal manera, entonces X y Y son dependientes.

Para variables aleatorias independientes continuas, también es cierto que la función de densidad conjunta $f(x, y)$ es el producto de la distribución de x sola, $f_1(x)$, y una función de y sola, $f_2(y)$, y éstas son las funciones de densidad (marginal) de X y Y respectivamente.

CAMBIO DE VARIABLES

Dadas las distribuciones de probabilidad de una o más variables aleatorias, con frecuencia nos interesamos en encontrar las distribuciones de otras variables aleatorias que dependen de ellas de manera específica. Los procedimientos para obtener estas distribuciones se presentan en los siguientes teoremas para el caso de variables discretas y continuas.

1. VARIABLES DISCRETAS.

Teorema 2-1: Sea X una variable aleatoria discreta cuya función de probabilidad es $f(x)$. Supongamos que la variable aleatoria discreta U se define en términos de X por $U = \phi(X)$, donde para cada valor de X corresponde uno y sólo uno de los valores de U e inversamente, de manera que $X = \psi(U)$. Entonces la función de probabilidad de U está dada por

$$g(u) = f[\psi(u)] \tag{31}$$

Teorema 2-2: Sean X y Y variables aleatorias discretas con una función de probabilidad conjunta $f(x, y)$. Supongamos que dos variables aleatorias discretas U y V están definidas en términos de X y Y por $U = \phi_1(X, Y)$, $V = \phi_2(X, Y)$, donde a cada par de valores de X y Y le corresponde uno y sólo un par de valores de U y V e inversamente, de manera que $X = \psi_1(U, V)$, $Y = \psi_2(U, V)$. Entonces la función de probabilidad conjunta de U y V está dada por

$$g(u, v) = f[\psi_1(u, v), \psi_2(u, v)] \tag{32}$$

2. VARIABLES CONTINUAS

Teorema 2-3: Sea X una variable aleatoria continua con probabilidad de densidad $f(x)$. Definamos $U = \phi(X)$ donde $X = \psi(U)$ como en el teorema 2-1. Entonces la densidad de probabilidad de U está dada por $g(u)$ donde

$$g(u)|du| = f(x)|dx| \tag{33}$$

o

$$g(u) = f(x)\left|\frac{dx}{du}\right| = f[\psi(u)]|\psi'(u)| \tag{34}$$

Teorema 2-4: Sean X y Y variables aleatorias continuas con función de densidad conjunta $f(x, y)$. Definamos $U = \phi_1(X, Y)$, $V = \phi_2(X, Y)$ donde $X = \psi_1(U, V)$, $Y = \psi_2(U, V)$ como en el teorema 2-2. Entonces la función de densidad conjunta de U y V está dada por $g(u, \psi)$ donde

$$g(u, v)|du\,dv| = f(x, y)|dx\,dy| \tag{35}$$

o

$$g(u, v) = f(x, y)\left|\frac{\partial(x, y)}{\partial(u, v)}\right| = f[\psi_1(u, v), \psi_2(u, v)]|J| \tag{36}$$

En (*36*) el determinante jacobiano, o simplemente jacobiano, está dado por

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} \tag{37}$$

DISTRIBUCIONES DE PROBABILIDAD DE FUNCIONES DE VARIABLES ALEATORIAS

Los teoremas 2-2 y 2-4 involucran específicamente funciones de probabilidad conjunta de dos variables aleatorias. En la práctica, con frecuencia se necesita encontrar la distribución de probabilidad de algunas funciones específicas de algunas variables aleatorias. Cualquiera de los siguientes teoremas es útil para tal propósito.

Teorema 2-5: Sean X y Y variables aleatorias continuas y sea $U = \phi_1(X, Y)$, $V = X$ (la segunda opción es arbitraria). Entonces, la función de densidad para U es la densidad marginal obtenida a partir de la densidad conjunta de U y V, tal como se encontró en el teorema 2-4. Un resultado similar es válido para las funciones de probabilidad de las variables discretas.

Teorema 2-6: Sea $f(x, y)$ la función de densidad conjunta de X y Y. Entonces la función de densidad $g(u)$ de la variable aleatoria $U = \phi_1(X, Y)$ se encuentra si se deriva con respecto a u la función de distribución dada por

$$G(u) = P[\phi_1(X, Y) \leqq u] = \iint_{\mathscr{R}} f(x, y)\, dx\, dy \tag{38}$$

donde $\mathscr{R}$ es la región para la cual $\phi_1(x, y) \leqq u$.

CONVOLUCIONES

Como consecuencia particular de los teoremas anteriores, podemos demostrar (*véase* problema 2.23) que la función de densidad de la suma de dos variables aleatorias continuas X y Y, es decir, de $U = X + Y$, con función de densidad $f(x, y)$ está dada por

$$g(u) = \int_{-\infty}^{\infty} f(x, u - x)\, dx \tag{39}$$

En el caso especial donde X y Y son independientes, $f(x, y) = f_1(x)f_2(y)$, y (*39*) se reduce a

$$g(u) = \int_{-\infty}^{\infty} f_1(x) f_2(u - x)\, dx \tag{40}$$

que recibe el nombre de *convolución* de f_1 y f_2, de manera abreviada $f_1 * f_2$.

Las siguientes son algunas propiedades importantes de la convolución:

1. $f_1 * f_2 = f_2 * f_1$
2. $f_1 * (f_2 * f_3) = (f_1 * f_2) * f_3$
3. $f_1 * (f_2 + f_3) = f_1 * f_2 + f_1 * f_3$

Estos resultados muestran que f_1, f_2, f_3 obedecen las leyes *conmutativa, asociativa y distributiva* del álgebra con respecto a las operaciones de la convolución.

DISTRIBUCIONES CONDICIONALES

Ya sabemos que si $P(A) > 0$,

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \tag{41}$$

Si X y Y son variables aleatorias discretas y tenemos los eventos $(A\colon X = x)$, $(B\colon Y = y)$, entonces (*41*) se convierte en

$$P(Y = y \mid X = x) = \frac{f(x, y)}{f_1(x)} \tag{42}$$

donde $f(x, y) = P(X = x, Y = y)$ es la función de probabilidad conjunta y $f_1(x)$ es la función de probabilidad marginal para X. Definimos

$$f(y \mid x) \equiv \frac{f(x, y)}{f_1(x)} \tag{43}$$

y la llamamos *función de probabilidad condicional de Y dado X*. De igual manera, la función de probabilidad condicional de X, dado Y, es

$$f(x \mid y) \equiv \frac{f(x, y)}{f_2(y)} \tag{44}$$

Algunas veces denotamos $f(x \mid y)$ y $f(y \mid x)$ por $f_1(x \mid y)$ y $f_2(y \mid x)$ respectivamente.

Estos conceptos se extienden fácilmente al caso donde X, Y son variables aleatorias continuas. Por ejemplo, *la función de densidad condicional de Y dado X* es

$$f(y \mid x) \equiv \frac{f(x, y)}{f_1(x)} \tag{45}$$

donde $f(x, y)$ es la función de densidad conjunta de X y Y, y $f_1(x)$ es la función de densidad marginal de X. Por ejemplo, usando (*45*) podemos encontrar la probabilidad de que Y se encuentre entre c y d dado que $x < X < x + dx$ es

$$P(c < Y < d \mid x < X < x + dx) = \int_c^d f(y \mid x)\, dy \tag{46}$$

Generalizaciones de estos resultados son también posibles.

APLICACIONES CON PROBABILIDAD GEOMÉTRICA

Varios problemas en probabilidad surgen de consideraciones geométricas o tienen interpretaciones geométricas. Por ejemplo, supongamos que tenemos un objetivo en forma de región plana de área K y una porción de ésta con área K_1, tal como aparece en la figura 2-5. Entonces, es razonable suponer que la probabilidad de

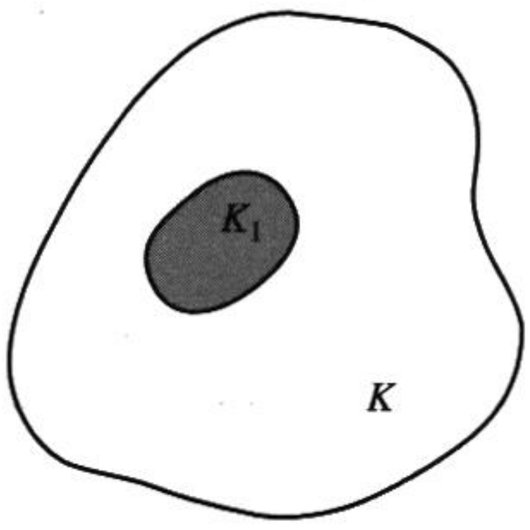


Fig. 2-5

acertar en la región de área K_1 es proporcional a K_1. Entonces definimos

$$P(\text{acertar en la región de área } K_1) = \frac{K_1}{K} \tag{47}$$

donde se supone que la probabilidad de dar en el blanco es 1. Clara que se pueden hacer otras suposiciones. Por ejemplo, podría haber menor probabilidad de acertar áreas exteriores. El tipo de suposición que se utiliza define la función de distribución de probabilidad.

Problemas resueltos

VARIABLES DISCRETAS ALEATORIAS Y DISTRIBUCIONES DE PROBABILIDAD

2.1. Suponga que se va a lanzar un par de dados balanceados y la variable aleatoria X denota la suma de los puntos. Obtenga la distribución de probabilidad para X.

Los puntos muestrales para los lanzamientos de un par de dados se muestran en la figura 1-9, página 16. La variable aleatoria X es la suma de las coordenadas para cada punto. Así para (3, 2) tenemos $X = 5$. Obtenemos la Tabla 2-4 teniendo en cuenta que los 36 puntos muestrales son igualmente probables, de manera que cada punto muestral tiene probabilidad de 1/36. Por ejemplo, correspondiendo a $x = 5$ tenemos los puntos muestrales (1, 4), (2, 3) (3, 2), (4, 1), de manera que la probabilidad asociada es 4/36.

Tabla 2-4

x	2	3	4	5	6	7	8	9	10	11	12
$f(x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

2.2. Encuentre la distribución de probabilidad de niños y niñas en familias con 3 hijos, suponiendo una probabilidad igual para niños y niñas.

El problema 1.37 trató el caso de n ensayos mutuamente independientes, en donde cada ensayo tenía dos resultados posibles, A y A', con las probabilidades respectivas p y $q = 1 - p$. Se encontró que la probabilidad de obtener exactamente x A en los n ensayos es ${}_nC_x p^x q^{n-x}$. Este resultado se aplica al problema presente, bajo la suposición de que los nacimientos (los ensayos) son independientes en lo que respecta al sexo del bebé. Así, siendo A el evento "varón", $n = 3$, y $p = q = \frac{1}{2}$, tenemos

$$P(\text{exactamente } x \text{ varones}) = P(X = x) = {}_3C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{3-x} = {}_3C_x \left(\frac{1}{2}\right)^3$$

donde la variable aleatoria X representa el número de hijos varones en la familia (observe que X se define para un espacio muestral de 3 ensayos). La probabilidad de la función para X,

$$f(x) = {}_3C_x \left(\frac{1}{2}\right)^3$$

se indica en la tabla 2-5.

Tabla 2-5

x	0	1	2	3
$f(x)$	1/8	3/8	3/8	1/8

FUNCIONES DE DISTRIBUCIÓN DISCRETA

2.3. (*a*) Encuentre la función de distribución $F(x)$ para la variable aleatoria X del problema 2.1, y (*b*) elabore una gráfica de esta función.

(*a*) Tenemos $F(x) = P(X \leqq x) = \sum_{u \leqq x} f(u)$. Entonces a partir de los resultados del problema 2.1, encontramos

$$F(x) = \begin{cases} 0 & -\infty < x < 2 \\ 1/36 & 2 \leqq x < 3 \\ 3/36 & 3 \leqq x < 4 \\ 6/36 & 4 \leqq x < 5 \\ \vdots & \vdots \\ 35/36 & 11 \leqq x < 12 \\ 1 & 12 \leqq x < \infty \end{cases}$$

(*b*) *Véase* figura 2-6.

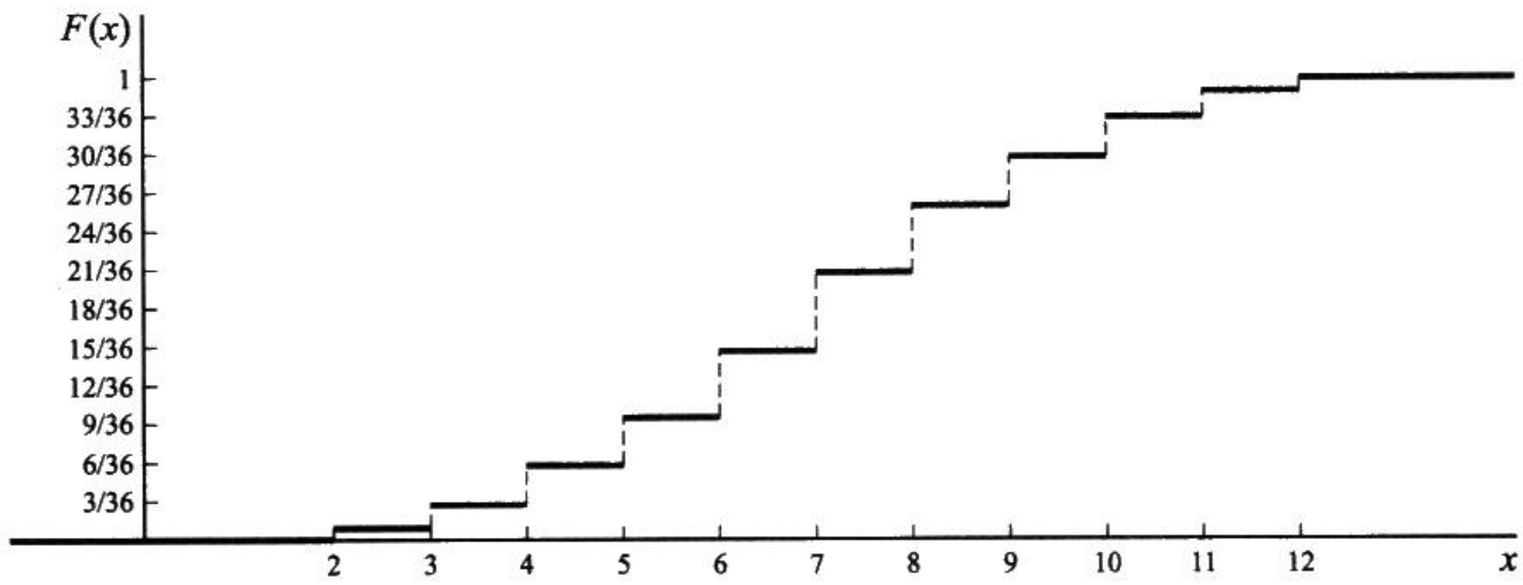


Fig. 2-6

2.4. (*a*) Encuentre la función de distribución $F(x)$ para la variable aleatoria X del problema 2.2, y (*b*) elabore una gráfica de esta función.

(*a*) Usando la tabla 2-5 del problema 2.2 obtenemos

$$F(x) = \begin{cases} 0 & -\infty < x < 0 \\ 1/8 & 0 \leqq x < 1 \\ 1/2 & 1 \leqq x < 2 \\ 7/8 & 2 \leqq x < 3 \\ 1 & 3 \leqq x < \infty \end{cases}$$

(*b*) La gráfica de la función de distribución de (*a*) se muestra en la figura 2-7.

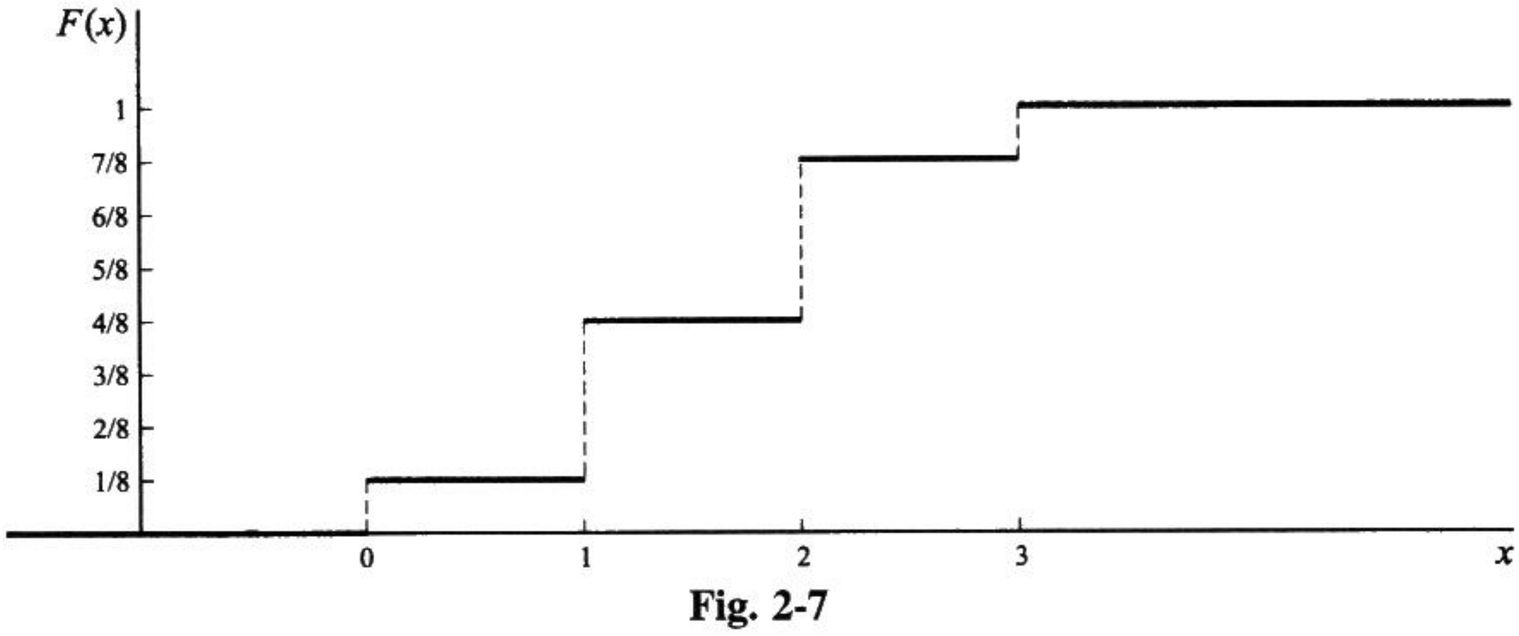


Fig. 2-7

VARIABLES ALEATORIAS CONTINUAS Y DISTRIBUCIONES DE PROBABILIDAD

2.5. Una variable aleatoria X tiene la función de densidad $f(x) = c/(x^2 + 1)$, donde $-\infty < x < \infty$. (a) Encuentre el valor de la constante c. (b) Encuentre la probabilidad de que X^2 se encuentre entre $1/3$ y 1.

(a) Debemos tener $\int_{-\infty}^{\infty} f(x)\,dx = 1$, es decir,

$$\int_{-\infty}^{\infty} \frac{c\,dx}{x^2+1} = c\tan^{-1} x\Big|_{-\infty}^{\infty} = c\left[\frac{\pi}{2} - \left(-\frac{\pi}{2}\right)\right] = 1$$

de manera que $c = 1/\pi$.

(b) Si $\frac{1}{3} \leqq X^2 \leqq 1$, entonces cualquier $\frac{\sqrt{3}}{3} \leqq X \leqq 1$ o $-1 \leqq X \leqq -\frac{\sqrt{3}}{3}$. Por lo tanto, la probabilidad requerida es

$$\frac{1}{\pi}\int_{-1}^{-\sqrt{3}/3} \frac{dx}{x^2+1} + \frac{1}{\pi}\int_{\sqrt{3}/3}^{1} \frac{dx}{x^2+1} = \frac{2}{\pi}\int_{\sqrt{3}/3}^{1} \frac{dx}{x^2+1}$$

$$= \frac{2}{\pi}\left[\tan^{-1}(1) - \tan^{-1}\left(\frac{\sqrt{3}}{3}\right)\right]$$

$$= \frac{2}{\pi}\left(\frac{\pi}{4} - \frac{\pi}{6}\right) = \frac{1}{6}$$

2.6. Encuentre la función de distribución correspondiente a la función de densidad del problema 2.5.

$$F(x) = \int_{-\infty}^{x} f(u)\,du = \frac{1}{\pi}\int_{-\infty}^{x} \frac{du}{u^2+1} = \frac{1}{\pi}\left[\tan^{-1} u\Big|_{-\infty}^{x}\right]$$

$$= \frac{1}{\pi}\left[\tan^{-1} x - \tan^{-1}(-\infty)\right] = \frac{1}{\pi}\left[\tan^{-1} x + \frac{\pi}{2}\right]$$

$$= \frac{1}{2} + \frac{1}{\pi}\tan^{-1} x$$

2.7. La función de distribución de una variable aleatoria X es

$$F(x) = \begin{cases} 1 - e^{-2x} & x \geqq 0 \\ 0 & x < 0 \end{cases}$$

Encuentre (a) la función de densidad, (b) la probabilidad de que $X > 2$, y (c) la probabilidad de que $-3 < X \leqq 4$.

(a) $$f(x) = \frac{d}{dx}F(x) = \begin{cases} 2e^{-2x} & x > 0 \\ 0 & x < 0 \end{cases}$$

(b) $$P(X > 2) = \int_{2}^{\infty} 2e^{-2u}\,du = -e^{-2u}\Big|_{2}^{\infty} = e^{-4}$$

Otro método.

Por definición, $P(X \leqq 2) = F(2) = 1 - e^{-4}$. Por lo tanto,

$$P(X > 2) = 1 - (1 - e^{-4}) = e^{-4}$$

(c) $$P(-3 < X \leqq 4) = \int_{-3}^{4} f(u)\,du = \int_{-3}^{0} 0\,du + \int_{0}^{4} 2e^{-2u}\,du$$

$$= -e^{-2u}\Big|_{0}^{4} = 1 - e^{-8}$$

Otro método.

$$P(-3 < X \leqq 4) = P(X \leqq 4) - P(X \leqq -3)$$
$$= F(4) - F(-3)$$
$$= (1 - e^{-8}) - (0) = 1 - e^{-8}$$

DISTRIBUCIONES CONJUNTAS Y VARIABLES INDEPENDIENTES

2.8. La función de probabilidad conjunta de dos variables aleatorias discretas X y Y está dada por $f(x, y) = c(2x + y)$, donde x y y pueden tomar todos los valores enteros tales que $0 \leqq x \leqq 2$, $0 \leqq y \leqq 3$, y $f(x, y) = 0$ de otra forma.

(*a*) Encuentre el valor de la constante c.

(*b*) Encuentre $P(X = 2, Y = 1)$.

(*c*) Encuentre $P(X \geqq 1, Y \leqq 2)$

(*a*) Los puntos muestrales (x, y) para los cuales las probabilidades son diferentes de cero se indican en la figura 2-8. Las probabilidades asociadas con estos puntos, dadas por $c(2x + y)$, se muestran en la tabla 2-6. Dado que el gran total, $42c$, debe ser igual a 1, tenemos $c = 1/42$.

Tabla 2-6

X \ Y	0	1	2	3	Totales ↓
0	0	c	$2c$	$3c$	$6c$
1	$2c$	$3c$	$4c$	$5c$	$14c$
2	$4c$	$5c$	$6c$	$7c$	$22c$
Totales→	$6c$	$9c$	$12c$	$15c$	$42c$

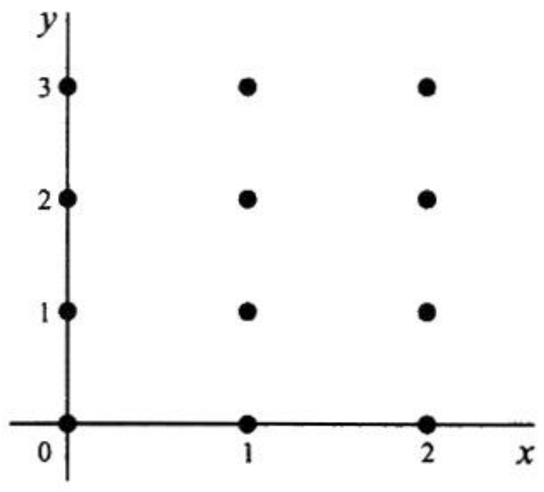


Fig. 2-8

(*b*) De la tabla 2-6 podemos ver que

$$P(X = 2,\ Y = 1) = 5c = \frac{5}{42}$$

(*c*) De la tabla 2-6 podemos ver que

$$P(X \geqq 1,\ Y \leqq 2) = \sum_{x \geqq 1} \sum_{y \leqq 2} f(x, y)$$
$$= (2c + 3c + 4c) + (4c + 5c + 6c)$$
$$= 24c = \frac{24}{42} = \frac{4}{7}$$

como lo indican las entradas sombreadas en la tabla.

2.9. Encuentre las funciones de probabilidad marginal (*a*) de X, y (*b*) de Y para las variables aleatorias del problema 2.8.

(*a*) La función de probabilidad marginal X está dada por $P(X = x) = f_1(x)$ y puede obtenerse de los totales del margen en la columna del extremo derecho de la tabla 2-6. De esto podemos ver que

$$P(X = x) = f_1(x) = \begin{cases} 6c = 1/7 & x = 0 \\ 14c = 1/3 & x = 1 \\ 22c = 11/21 & x = 2 \end{cases}$$

Verificación: $\frac{1}{7} + \frac{1}{3} + \frac{11}{21} = 1$

(*b*) La función de probabilidad marginal para Y está dada por $P(Y = y) = f_2(y)$ y puede obtenerse de los totales del margen de la última fila de la tabla 2-6. De esto podemos ver que

$$P(Y = y) = f_2(y) = \begin{cases} 6c = 1/7 & y = 0 \\ 9c = 3/14 & y = 1 \\ 12c = 2/7 & y = 2 \\ 15c = 5/14 & y = 3 \end{cases}$$

Verificación: $\frac{1}{7} + \frac{3}{14} + \frac{2}{7} + \frac{5}{14} = 1$

2.10. Demuestre que las variables aleatorias X y Y del problema 2.8 son dependientes.

Si las variables X y Y son independientes, deben tener, para todo x y y,

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

Pero como se vio en los problemas 2.8(*b*) y 2.9,

$$P(X = 2,\ Y = 1) = \frac{5}{42} \quad ; \qquad P(X = 2) = \frac{11}{21} \quad ; \qquad P(Y = 1) = \frac{3}{14}$$

de manera que $$P(X = 2, Y = 1) \neq P(X = 2)P(Y = 1)$$

El resultado también se deduce del hecho de que la función de probabilidad conjunta $(2x + y)/42$ no puede expresarse como función de x sola por la función de y sola.

2.11. La función de densidad conjunta de dos variables aleatorias continuas X y Y es

$$f(x, y) = \begin{cases} cxy & 0 < x < 4,\ 1 < y < 5 \\ 0 & \text{de otra manera} \end{cases}$$

(*a*) Encuentre el valor de la constante c.

(*b*) Encuentre $P(1 < X < 2, 2 < Y < 3)$.

(*c*) Encuentre $P(X \geqq 3, Y \leqq 2)$.

(*a*) Debemos igualar a 1 la probabilidad total, es decir,

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$$

Usando la definición de $f(x, y)$, la integral tiene el valor

$$\int_{x=0}^{4} \int_{y=1}^{5} cxy\, dx\, dy = c \int_{x=0}^{4} \left[\int_{y=1}^{5} xy\, dy \right] dx$$

$$= c \int_{z=0}^{4} \frac{xy^2}{2} \Big|_{y=1}^{5} dx = c \int_{x=0}^{4} \left(\frac{25x}{2} - \frac{x}{2} \right) dx$$

$$= c \int_{x=0}^{4} 12x\, dx = c(6x^2) \Big|_{x=0}^{4} = 96c$$

Entonces $96c = 1$ y $c = 1/96$.

(*b*) Usando el valor de c encontrado en (*a*), tenemos

$$P(1 < X < 2, 2 < Y < 3) = \int_{x=1}^{2} \int_{y=2}^{3} \frac{xy}{96}\, dx\, dy$$

$$= \frac{1}{96} \int_{x=1}^{2} \left[\int_{y=2}^{3} xy\, dy \right] dx = \frac{1}{96} \int_{x=1}^{2} \frac{xy^2}{2} \Big|_{y=2}^{3} dx$$

$$= \frac{1}{96} \int_{x=1}^{2} \frac{5x}{2}\, dx = \frac{5}{192} \left(\frac{x^2}{2} \right) \Big|_{1}^{2} = \frac{5}{128}$$

(*c*)
$$P(X \geqq 3,\ Y \leqq 2) = \int_{x=3}^{4} \int_{y=1}^{2} \frac{xy}{96}\, dx\, dy$$
$$= \frac{1}{96} \int_{x=3}^{4} \left[\int_{y=1}^{2} xy\, dy \right] dx = \frac{1}{96} \int_{x=3}^{4} \frac{xy^2}{2} \Big|_{y=1}^{2} dx$$
$$= \frac{1}{96} \int_{x=3}^{4} \frac{3x}{2}\, dx = \frac{7}{128}$$

2.12. Encuentre las funciones de distribución marginal (*a*) de X y (*b*) de Y para el problema 2.11.

(*a*) La función de distribución marginal para X si $0 \leqq x \leqq 4$ es

$$F_1(x) = P(X \leqq x) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{\infty} f(u, v)\, du\, dv$$
$$= \int_{u=0}^{x} \int_{v=1}^{5} \frac{uv}{96}\, du\, dv$$
$$= \frac{1}{96} \int_{u=0}^{x} \left[\int_{v=1}^{5} uv\, dv \right] du = \frac{x^2}{16}$$

Para $X \geqq 4$, $F_1(x) = 1$; para $x < 0$, $F_1(x) = 0$. Por lo tanto,

$$F_1(x) = \begin{cases} 0 & x < 0 \\ x^2/16 & 0 \leqq x < 4 \\ 1 & x \geqq 4 \end{cases}$$

Dado que $F_1(x)$ es continua entre $x = 0$ y $x = 4$, podemos remplazar $<$ por $\leqq$ en las expresión anterior.

(*b*) La función de distribución marginal para Y, si $1 \leqq y < 5$, es

$$F_2(y) = P(Y \leqq y) = \int_{u=-\infty}^{\infty} \int_{v=1}^{y} f(u, v)\, du\, dv$$
$$= \int_{u=0}^{4} \int_{v=1}^{y} \frac{uv}{96}\, du\, dv = \frac{y^2 - 1}{24}$$

Para $y \geqq 5$, $F_2(y) = 1$. Para $y < 1$, $F_2(y) = 0$. Así que

$$F_2(y) = \begin{cases} 0 & y < 1 \\ (y^2 - 1)/24 & 1 \leqq y < 5 \\ 1 & y \geqq 5 \end{cases}$$

Dado que $F_2(y)$ es continua entre $y = 1$ y $y = 5$, podemos remplazar $<$ por $\leqq$ en la expresión anterior.

2.13. Encuentre la función de distribución conjunta para las variables aleatorias X, Y del problema 2.11.

Del problema 2.11 vemos que la función de densidad conjunta para X y Y se puede escribir como el producto de una función de x sola y una función de y sola. De hecho, $f(x, y) = f_1(x)f_2(y)$, donde

$$f_1(x) = \begin{cases} c_1 x & 0 < x < 4 \\ 0 & \text{de otra manera} \end{cases} \qquad f_2(y) = \begin{cases} c_2 y & 1 < y < 5 \\ 0 & \text{de otra manera} \end{cases}$$

y $c_1 c_2 = c = 1/96$. Se deduce que X y Y son independientes, de manera que su función de distribución conjunta está dada por $F(x, y) = F_1(x)F_2(y)$. Las distribuciones marginales $F_1(x)$ y $F_2(y)$ se determinaron en el problema 2.12, y la figura 2-9 muestra la distribución por trazos de $F(x, y)$.

2.14. Encuentre $P(X + Y < 3)$ en el problema 2.11.

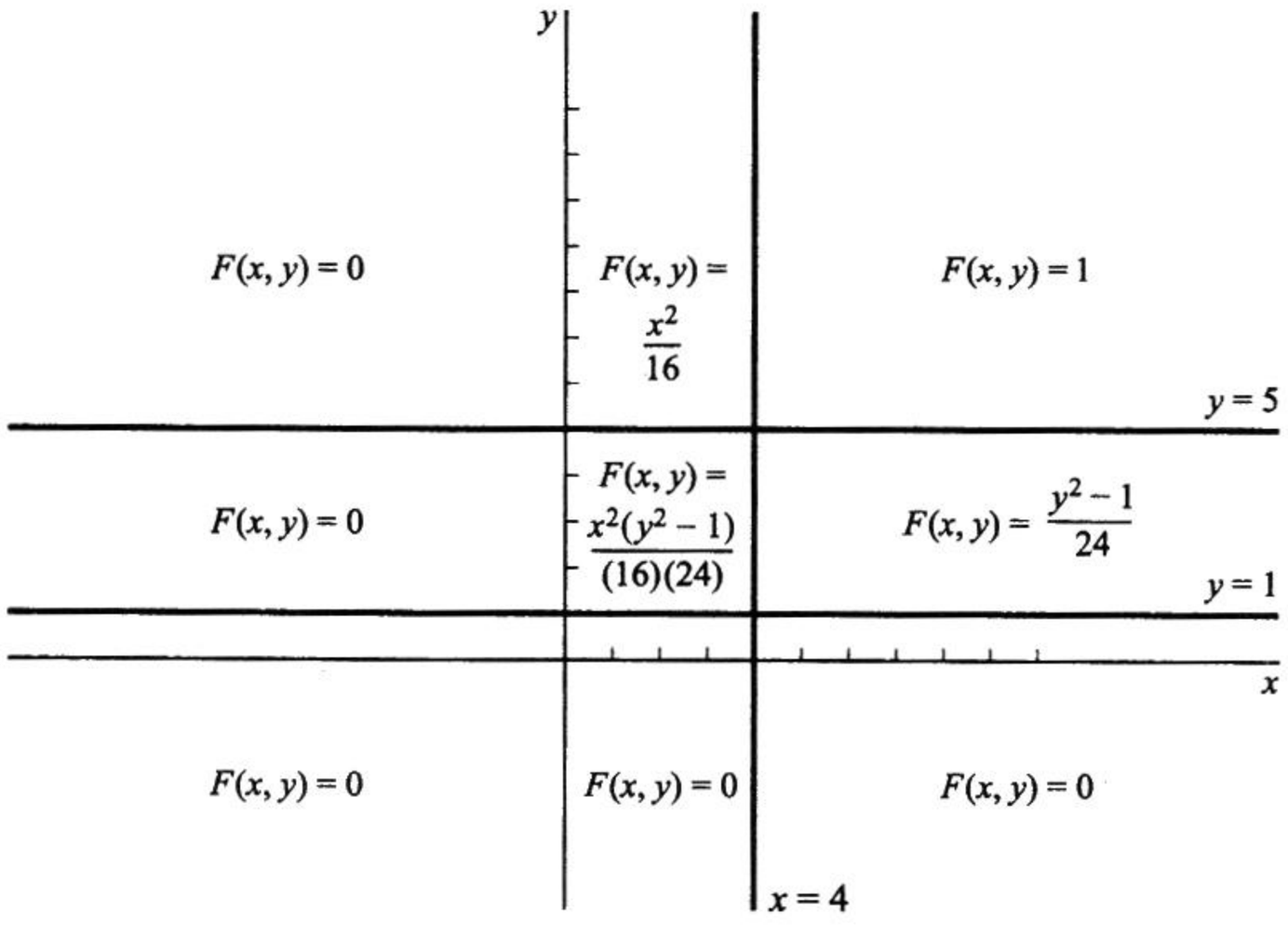


Fig. 2-9

En la figura 2-10 indicamos que la región cuadrada $0 < x < 4$, $1 < y < 5$ dentro de la cual la función de distribución de X y Y es diferente de cero. La probabilidad requerida está dada por

$$P(X + Y < 3) = \iint_{\mathscr{R}} f(x, y)\, dx\, dy$$

donde $\mathscr{R}$ es la parte del cuadrado sobre la cual $x + y < 3$ se ve sombreado en la figura 2-10. Dado que $f(x, y) = xy/96$ sobre $\mathscr{R}$, esta probabilidad está dada por

$$\int_{x=0}^{2} \int_{y=1}^{3-x} \frac{xy}{96}\, dx\, dy$$

$$= \frac{1}{96} \int_{x=0}^{2} \left[\int_{y=1}^{3-x} xy\, dy \right] dx$$

$$= \frac{1}{96} \int_{x=0}^{2} \frac{xy^2}{2} \Big|_{y=1}^{3-x} dx = \frac{1}{192} \int_{x=0}^{2} [x(3 - x)^2 - x] = \frac{1}{48}$$

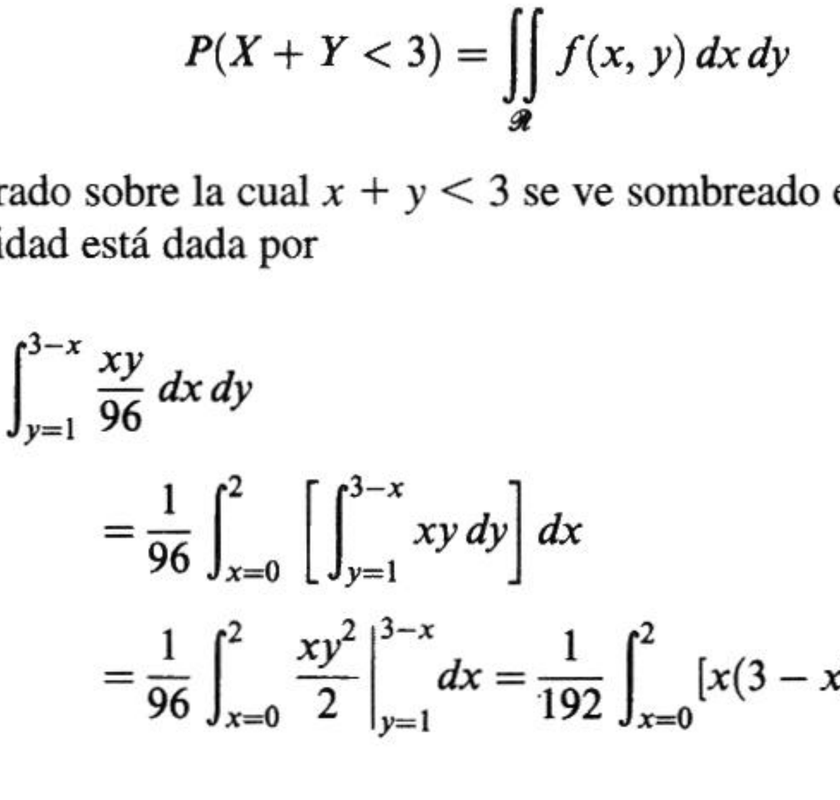


Fig. 2-10

CAMBIO DE VARIABLES

2.15. Demuestre el teorema 2-1, página 47.

La función de probabilidad para U está dada por

$$g(u) = P(U = u) = P[\phi(X) = u] = P[X = \psi(u)] = f[\psi(u)]$$

De manera similar podemos demostrar el teorema 2-2, página 47.

2.16. Demuestre el teorema 2-3, página 47.

Considere primero el caso donde $u = \phi(x)$ o $x = \psi(u)$ es una función creciente, es decir, u se incrementa a medida que x se incrementa (figura 2-11). Entonces, como podemos deducir de la figura, tenemos

(*1*) $$P(u_1 < U < u_2) = P(x_1 < X < x_2)$$

o

(*2*) $$\int_{u_1}^{u_2} g(u)\,du = \int_{x_1}^{x_2} f(x)\,dx$$

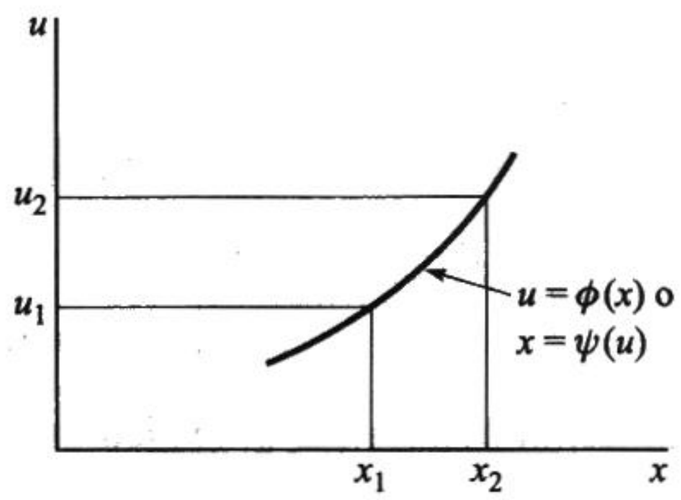


Fig. 2-11

Sea $x = \psi(u)$ en la integral de la derecha, podemos escribir (*2*)

$$\int_{u_1}^{u_2} g(u)\,du = \int_{u_1}^{u_2} f[\psi(u)]\psi'(u)\,du$$

Esto es válido para todo u_1 y u_2 sólo si las integrales son idénticas, es decir,

$$g(u) = f[\psi(u)]\psi'(u)$$

Éste es un caso especial de (*34*), página 47, donde $\psi'(u) > 0$ (es decir que la pendiente es positiva). Para el caso en donde $\psi'(u) \leqq 0$, es decir, u es función decreciente de x, podemos demostrar también que (*34*) es válido (*véase* problema 2.67). El teorema también puede demostrarse si $\psi'(u) \geqq 0$ o $\psi'(u) < 0$.

2.17. Demuestre el teorema 2-4, página 47.

Primero suponemos que x y y se incrementan, y también lo hacen u y v. Tal como en el problema 2.16 podemos demostrar que

$$P(u_1 < U < u_2,\ v_1 < V < v_2) = P(x_1 < X < x_2,\ y_1 < Y < y_2)$$

o

$$\int_{u_1}^{u_2}\int_{v_1}^{v_2} g(u, v)\,du\,dv = \int_{x_1}^{x_2}\int_{y_1}^{y_2} f(x, y)\,dx\,dy$$

Si en la integral de la derecha $x = \psi_1(u, v)$, y $= \psi_2(u, v)$, tenemos, de un teorema de cálculo avanzado,

$$\int_{v_1}^{u_2}\int_{v_1}^{v_2} g(u, v)\,du\,dv = \int_{u_1}^{u_2}\int_{v_1}^{v_2} f[\psi_1(u, v), \psi_2(u, v)]J\,du\,dv$$

donde

$$J = \frac{\partial(x, y)}{\partial(u, v)}$$

es el jacobiano. Así que

$$g(u, v) = f[\psi_1(u, v), \psi_2(u, v)]J$$

la cual es (*36*), página 47, en el caso donde $J > 0$. De igual modo, podemos demostrar (*36*) para el caso $J < 0$.

2.18. La función de probabilidad de una variable aleatoria X es

$$f(x) = \begin{cases} 2^{-x} & x = 1, 2, 3, \ldots \\ 0 & \text{de otra manera} \end{cases}$$

Encuentre la función de probabilidad para la variable aleatoria $U = X^4 + 1$.

Dado que $U = X^4 + 1$, la relación entre los valores u y x de las variables aleatorias U y X está dada por $u = x^4 + 1$ o $x = \sqrt[4]{u - 1}$, donde $u = 2, 17, 82, \ldots$ y se toma la raíz real positiva. Entonces, la función de probabilidad requerida para U está dada por

$$g(u) = \begin{cases} 2^{-\sqrt[4]{u-1}} & u = 2, 17, 82, \ldots \\ 0 & \text{de otra manera} \end{cases}$$

usando el teorema 2-1, página 47, o el problema 2.15.

2.19. La función de probabilidad de una variable aleatoria X está dada por

$$f(x) = \begin{cases} x^2/81 & -3 < x < 6 \\ 0 & \text{de otra manera} \end{cases}$$

Encuentre la densidad de probabilidad para la variable aleatoria $U = (\frac{1}{3})(12 - X)$.

Tenemos que $u = (\frac{1}{3})(12 - x)$ o $x = 12 - 3u$. Así, para cada valor de x hay uno y sólo un valor de u e inversamente. Los valores de u correspondientes a $x = -3$ y $x = 6$ son $u = 5$ y $u = 2$, respectivamente. Dado que $\psi'(u) = dx/du = -3$, se deduce por el teorema 2-3, página 47, o el problema 2.16 que la función de densidad para U es

$$g(u) = \begin{cases} (12 - 3u)^2/27 & 2 < u < 5 \\ 0 & \text{de otra manera} \end{cases}$$

Verificación:

$$\int_2^5 \frac{(12 - 3u)^2}{27}\, du = -\frac{(12 - 3u)^3}{243}\bigg|_2^5 = 1$$

2.20. Encuentre la densidad de probabilidad de la variable aleatoria $U = X^2$ donde X es la variable aleatoria del problema 2.19.

Tenemos que $u = x^2$ o $x = \pm\sqrt{u}$. Así, para cada valor de x le corresponde uno y sólo un valor de u, pero a cada valor de $u \neq 0$ le corresponden *dos* valores de x. Los valores de x para los cuales $-3 < x < 6$ corresponden a valores de u para los cuales $0 \leqq u < 36$, como aparece en la figura 2-12.

Como se ve en esta figura, el intervalo $-3 < x \leqq 3$ corresponde a $0 \leqq u \leqq 9$ mientras que $3 < x < 6$ corresponde a $9 < u < 36$. En este caso no podemos usar directamente el teorema 2-3, pero podemos proceder de la siguiente manera. La función de distribución para U es

$$G(u) = P(U \leqq u)$$

Ahora, si $0 \leqq u \leqq 9$, tenemos

$$G(u) = P(U \leqq u) = P(X^2 \leqq u) = P(-\sqrt{u} \leqq X \leqq \sqrt{u})$$

$$= \int_{-\sqrt{u}}^{\sqrt{u}} f(x)\, dx$$

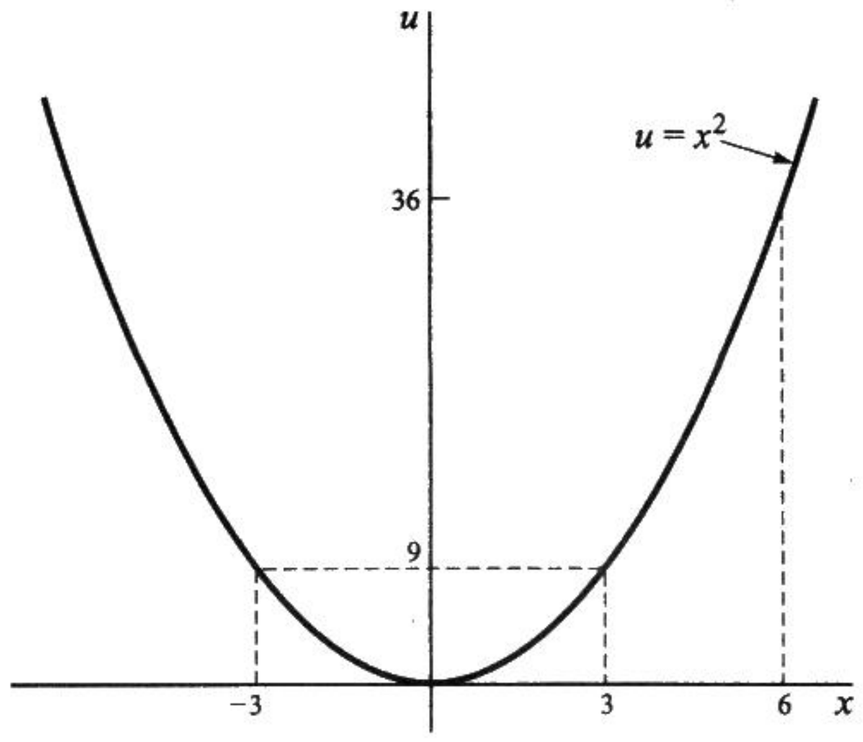


Fig. 2-12

Pero si $9 < u < 36$, tenemos

$$G(u) = P(U \leqq u) = P(-3 < X < \sqrt{u}) = \int_{-3}^{\sqrt{u}} f(x)\,dx$$

Dado que la función de densidad $g(u)$ es la derivada de $G(u)$, tenemos que usando (*12*),

$$g(u) = \begin{cases} \dfrac{f(\sqrt{u}) + f(-\sqrt{u})}{2\sqrt{u}} & 0 \leqq u \leqq 9 \\ \dfrac{f(\sqrt{u})}{2\sqrt{u}} & 9 < u < 36 \\ 0 & \text{de otra manera} \end{cases}$$

Usando la definición dada de $f(x)$, esto se convierte en

$$g(u) = \begin{cases} \sqrt{u}/81 & 0 \leqq u \leqq 9 \\ \sqrt{u}/162 & 9 < u < 36 \\ 0 & \text{de otra manera} \end{cases}$$

Verificación:

$$\int_0^9 \frac{\sqrt{u}}{81}\,du + \int_9^{36} \frac{\sqrt{u}}{162}\,du = \left.\frac{2u^{3/2}}{243}\right|_0^9 + \left.\frac{u^{3/2}}{243}\right|_9^{36} = 1$$

2.21. Si las variables aleatorias X y Y tienen función de densidad conjunta

$$f(x, y) = \begin{cases} xy/96 & 0 < x < 4,\ 1 < y < 5 \\ 0 & \text{de otra manera} \end{cases}$$

(*véase* problema 2.11), encuentre la función de densidad de $U = X + 2Y$.

Método 1.

Sea $u = x + 2y$, $v = x$, escogiendo la segunda relación arbitrariamente. Entonces la solución simultánea da como resultado $x = v$, $y = (\frac{1}{2})(u - v)$. Así la región $0 < x < 4$, $1 < y < 5$ corresponde a la región $0 < v < 4$, $2 < u - v < 10$, mostradas en la parte sombreada de la figura 2-13.

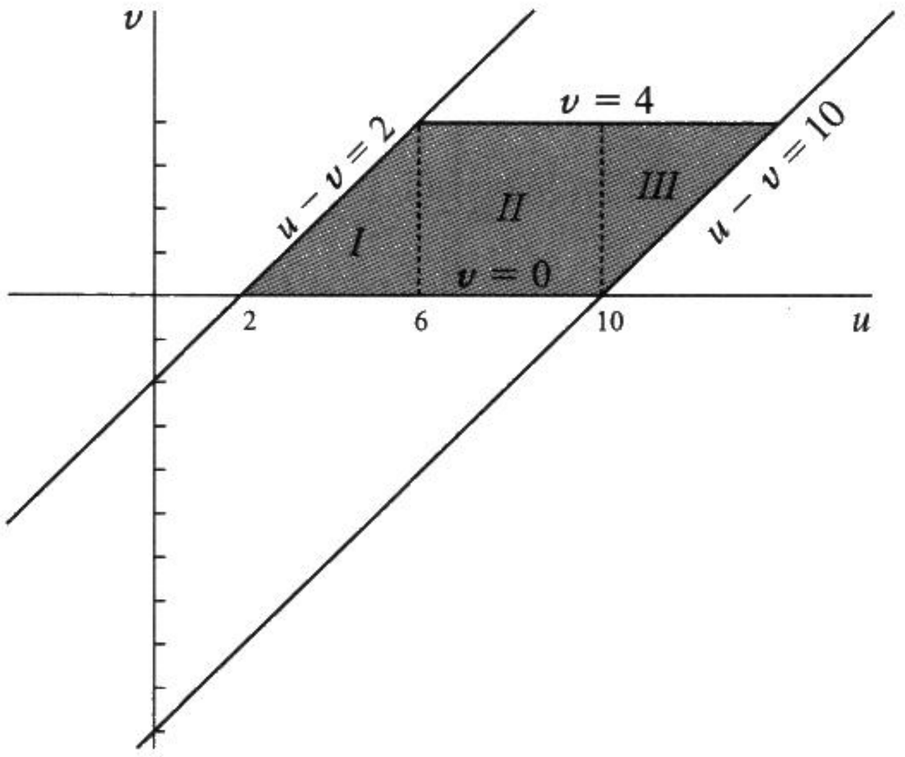


Fig. 2-13

el jacobiano está dado por

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}$$

Entonces según el teorema 2-4 la función de densidad conjunta de U y V es

$$g(u, v) = \begin{cases} v(u-v)/384 & 2 < u - v < 10, \quad 0 < v < 4 \\ 0 & \text{de otra manera} \end{cases}$$

La función de densidad marginal de U está dada por

$$g_1(u) = \begin{cases} \displaystyle\int_{v=0}^{u-2} \frac{v(u-v)}{384}\, dv & 2 < u < 6 \\ \displaystyle\int_{v=0}^{4} \frac{v(u-v)}{384}\, dv & 6 < u < 10 \\ \displaystyle\int_{v=u-10}^{4} \frac{v(u-v)}{384}\, dv & 10 < u < 14 \\ 0 & \text{de otra manera} \end{cases}$$

como se observa al referirnos a las regiones sombreadas *I*, *II*, *III* de la figura 2-13. Desarrollando las integrales encontramos,

$$g_1(u) = \begin{cases} (u-2)^2(u+4)/2304 & 2 < u < 6 \\ (3u-8)/144 & 6 < u < 10 \\ (348u - u^3 - 2128)/2304 & 10 < u < 14 \\ 0 & \text{de otra manera} \end{cases}$$

Se puede llevar a cabo una verificación demostrando que la integral de $g_1(u)$ es igual a 1.

Método 2.

La función de distribución de la variable aleatoria $X + 2Y$ está dada por

$$P(X + 2Y \leqq u) = \iint_{x+2y \leqq u} f(x, y)\, dx\, dy = \iint_{\substack{x+2y \leqq u \\ 0<x<4 \\ 1<y<5}} \frac{xy}{96}\, dx\, dy$$

Para $2 < u < 6$ vemos, refiriéndonos a la figura 2-14, que la última integral es igual a

$$\int_{x=0}^{u-2} \int_{y=1}^{(u-x)/2} \frac{xy}{96}\, dx\, dy = \int_{x=0}^{u-2} \left[\frac{x(u-x)^2}{768} - \frac{x}{192} \right] dx$$

Al derivar esta integral con respecto a u, tenemos $(u - 2)^2(u + 4)/2.304$. De igual manera podemos obtener el resultado del método 1 para $6 < u < 10$, etc.

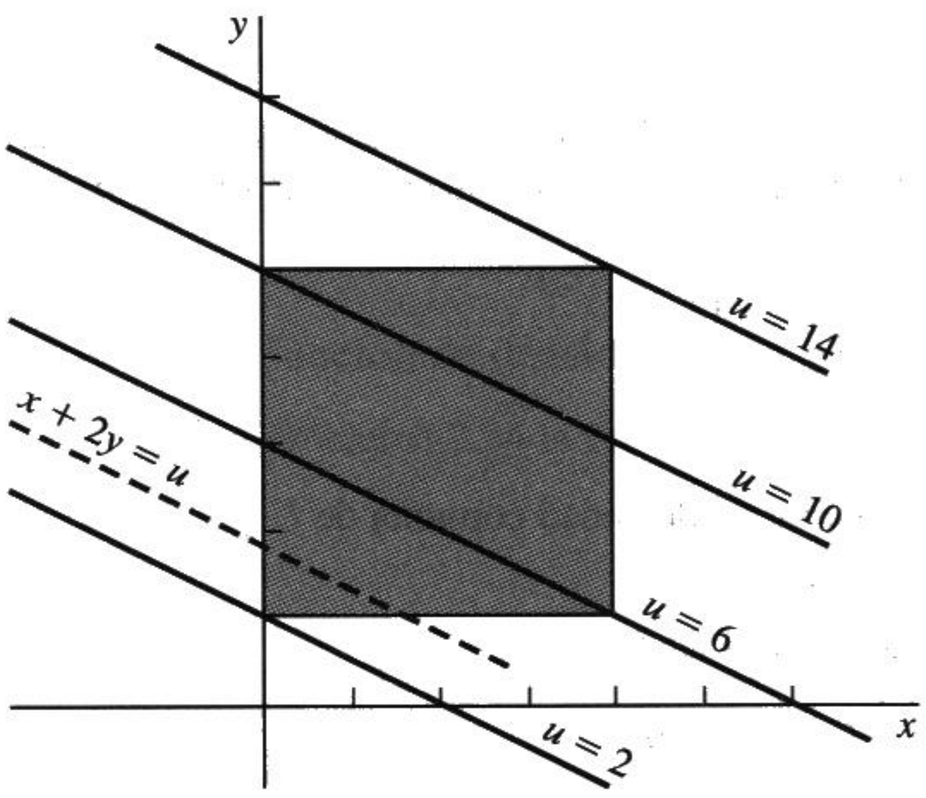


Fig. 2-14

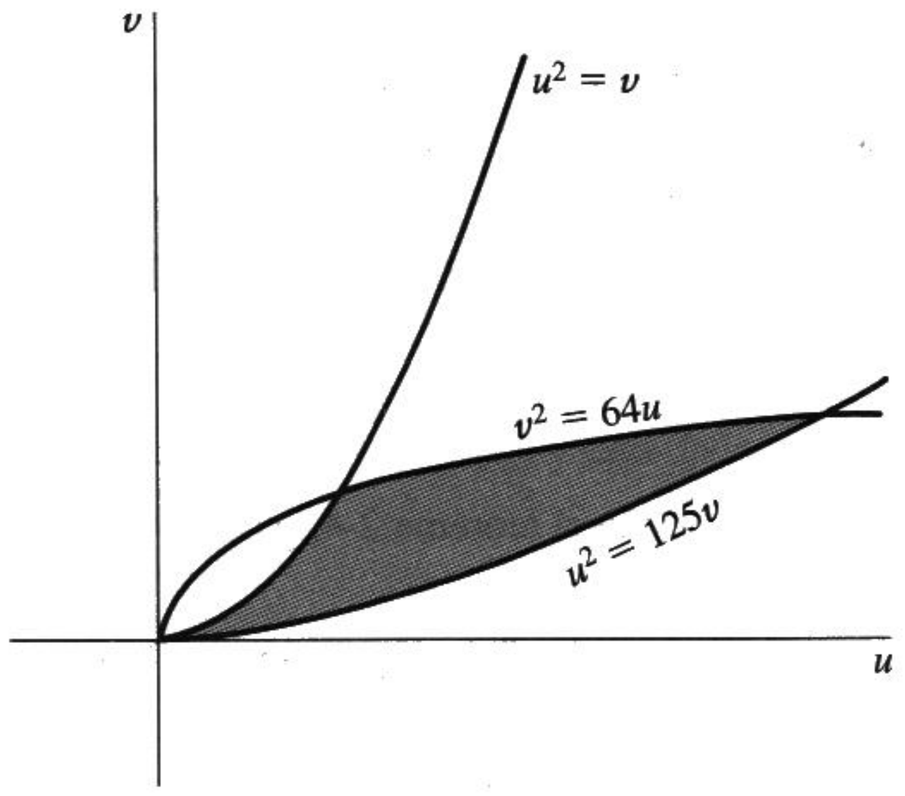


Fig. 2-15

2.22. Si las variables aleatorias X y Y tienen función de densidad conjunta

$$f(x, y) = \begin{cases} xy/96 & 0 < x < 4,\ 1 < y < 5 \\ 0 & \text{de otra manera} \end{cases}$$

(*véase* el problema 2.11), encuentre la función de densidad conjunta de $U = XY^2$, $V = X^2Y$.

Consideremos $u = xy^2$, $v = x^2y$. Dividiendo estas ecuaciones obtenemos $y/x = u/v$ de manera que $y = ux/v$. Esto no lleva a la solución simultánea $x = v^{2/3}u^{-1/3}$, $y = u^{2/3}v^{-1/3}$. La imagen de $0 < x < 4$, $1 < y < 5$ en el plano uv está dada por

$$0 < v^{2/3}u^{-1/3} < 4 \quad ; \quad 1 < u^{2/3}v^{-1/3} < 5$$

lo cual equivale a

$$v^2 < 64u \quad ; \quad v < u^2 < 125v$$

Esta región aparece sombreada en la figura 2-15.

El jacobiano está dado por

$$J = \begin{vmatrix} -\frac{1}{3}\, v^{2/3}u^{-4/3} & \frac{2}{3}\, v^{-1/3}u^{-1/3} \\ \frac{2}{3}\, u^{-1/3}v^{-1/3} & -\frac{1}{3}\, u^{2/3}v^{-4/3} \end{vmatrix} = -\frac{1}{3}\, u^{-2/3}v^{-2/3}$$

Así la función de densidad conjunta de U y V es, según el teorema 2-4

$$g(u, v) = \begin{cases} \dfrac{(v^{2/3}u^{-1/3})(u^{2/3}v^{-1/3})}{96}\left(\tfrac{1}{3}u^{-2/3}v^{-2/3}\right) & v^2 < 64u,\ v < u^2 < 125v \\ 0 & \text{de otra manera} \end{cases}$$

o

$$g(u, v) = \begin{cases} u^{-1/3}v^{-1/3}/288 & v^2 < 64u, \quad v < u^2 < 125v \\ 0 & \text{de otra manera} \end{cases}$$

CONVOLUCIONES

2.23. Sean X y Y variables aleatorias con función de densidad conjunta $f(x, y)$. Demuestre que la función de densidad de $U = X + Y$ es

$$g(u) = \int_{-\infty}^{\infty} f(v, u - v)\, dv$$

Método 1.

Sea $U = X + Y, V = X$, donde hemos añadido arbitrariamente la segunda ecuación. A estas ecuaciones les corresponde $u = x + y, v = x$ o $x = v, y = u - v$. El jacobiano de la transformación está dado por

$$J = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ 1 & -1 \end{vmatrix} = -1$$

Así, según el teorema 2-4, página 47, tenemos que la función de densidad conjunta de U y V es

$$g(u, v) = f(v, u - v)$$

Se deduce de (*26*), página 46, que la función de densidad marginal de U es

$$g(u) = \int_{-\infty}^{\infty} f(v, u - v)\, dv$$

Método 2.

La función de distribución de $U = X + Y$ es igual a la integral doble de $f(x, y)$ tomada sobre la región definida por $x + y \leqq u$, es decir,

$$G(u) = \iint_{x+y \leqq u} f(x, y)\, dx\, dy$$

Dado que dicha región está por debajo de la línea $x + y = u$, como lo indica la parte sombreada de la figura 2-16, vemos que

$$G(u) = \int_{x=-\infty}^{\infty} \left[\int_{y=-\infty}^{u-x} f(x, y)\, dy \right] dx$$

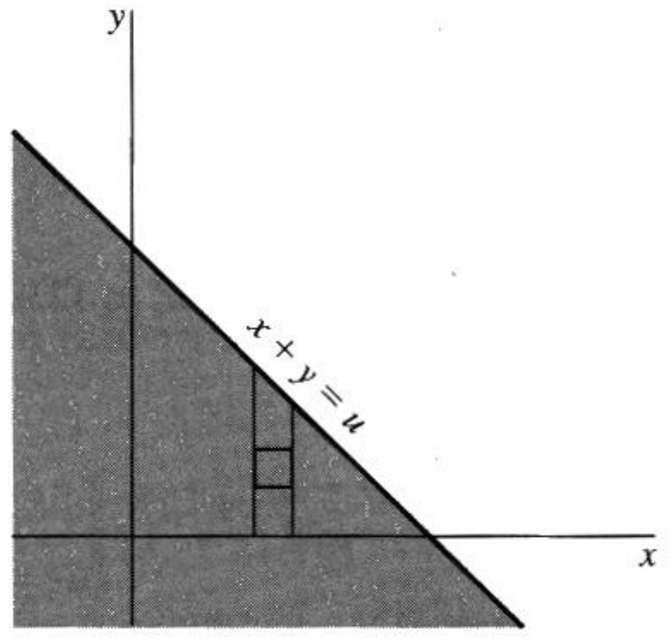


Fig. 2-16

La función de densidad de U es la derivada de $G(u)$ con respecto a u, y está dada por

$$g(u) = \int_{-\infty}^{\infty} f(x, u - x)\, dx$$

usando primero (*12*) en la integral de x y luego en la integral de y.

2.24. Desarrolle el problema 2.23 si X y Y son variables aleatorias independientes con funciones de densidad $f_1(x)$ y $f_2(y)$, respectivamente.

En este caso la función de densidad conjunta es $f(x, y) = f_1(x)\, f_2(y)$, de manera que según el problema 2.23 la función de densidad de $U = X + Y$ es

$$g(u) = \int_{-\infty}^{\infty} f_1(v) f_2(u - v)\, dv = f_1 * f_2$$

la cual es la *convolución* de f_1 y f_2.

2.25. Si X y Y son variables aleatorias independientes con funciones de densidad

$$f_1(x) = \begin{cases} 2e^{-2x} & x \geqq 0 \\ 0 & x < 0 \end{cases} \quad ; \quad f_2(y) = \begin{cases} 3e^{-3y} & y \geqq 0 \\ 0 & y < 0 \end{cases}$$

encuentre la función de densidad de su suma, $U = X + Y$.

Según el problema 2.24 la función de densidad requerida es la convolución de f_1 y f_2 y está dada por

$$g(u) = f_1 * f_2 = \int_{-\infty}^{\infty} f_1(v) f_2(u - v)\, dv$$

En el integrando f_1 se anula cuando $v < 0$ y f_2 se anula cuando $v > u$. Por lo tanto,

$$g(u) = \int_0^u (2e^{-2v})(3e^{-3(u-v)})\, dv$$

$$= 6e^{-3u} \int_0^u e^v\, dv = 6e^{-3u}(e^u - 1) = 6(e^{-2u} - e^{-3u})$$

si $u \geqq 0$ y $g(u) = 0$ si $u < 0$.

Verificación: $$\int_{-\infty}^{\infty} g(u)\, du = 6 \int_0^{\infty} (e^{-2u} - e^{-3u})\, du = 6\left(\frac{1}{2} - \frac{1}{3}\right) = 1$$

2.26. Demuestre que $f_1 * f_2 = f_2 * f_1$ (propiedad 1, página 48).

Tenemos

$$f_1 * f_2 = \int_{v=-\infty}^{\infty} f_1(v) f_2(u - v)\, dv$$

Sea $w = u - v$ de manera que $v = u - w$, $dv = -dw$, obtenemos

$$f_1 * f_2 = \int_{w=\infty}^{-\infty} f_1(u - w) f_2(w)(-dw) = \int_{w=-\infty}^{\infty} f_2(w) f_1(u - w)\, dw = f_2 * f_1$$

DISTRIBUCIONES CONDICIONALES

2.27. Encuentre (*a*) $f(y \mid 2)$, (*b*) $P(Y = 1 \mid X = 2)$ para la distribución del problema 2.8.

(*a*) Usando los resultados de los problemas 2.8 y 2.9, tenemos

$$f(y \mid x) = \frac{f(x, y)}{f_1(x)} = \frac{(2x + y)/42}{f_1(x)}$$

de manera que $x = 2$

$$f(y \mid 2) = \frac{(4 + y)/42}{11/21} = \frac{4 + y}{22}$$

(*b*) $$P(Y = 1 \mid X = 2) = f(1 \mid 2) = \frac{5}{22}$$

2.28. Si X y Y tienen función de densidad conjunta

$$f(x, y) = \begin{cases} \frac{3}{4} + xy & 0 < x < 1, 0 < y < 1 \\ 0 & \text{de otra manera} \end{cases}$$

encuentre (*a*) $f(y \mid x)$, (*b*) $P(Y > \frac{1}{2} \mid \frac{1}{2} < X < \frac{1}{2} + dx)$.

(*a*) Para $0 < x < 1$,

$$f_1(x) = \int_0^1 \left(\frac{3}{4} + xy\right) dy = \frac{3}{4} + \frac{x}{2}$$

y

$$f(y \mid x) = \frac{f(x, y)}{f_1(x)} = \begin{cases} \dfrac{3 + 4xy}{3 + 2x} & 0 < y < 1 \\ 0 & \text{otro } y \end{cases}$$

Para otros valores de x, $f(y \mid x)$ no está definida.

(*b*) $$P(Y > \tfrac{1}{2} \mid \tfrac{1}{2} < X < \tfrac{1}{2} + dx) = \int_{1/2}^{\infty} f(y \mid \tfrac{1}{2})\, dy = \int_{1/2}^{1} \frac{3 + 2y}{4}\, dy = \frac{9}{16}$$

2.29. La función de densidad conjunta de las variables aleatorias X y Y está dada por

$$f(x, y) = \begin{cases} 8xy & 0 \leqq x \leqq 1, 0 \leqq y \leqq x \\ 0 & \text{de otra manera} \end{cases}$$

Encuentre (*a*) la densidad marginal de X, (*b*) la densidad marginal de Y, (*c*) la densidad condicional de X, (*d*) la densidad condicional de Y.

La región sobre la cual $f(x, y)$ es diferente de cero aparece sombreada en la figura 2-17.

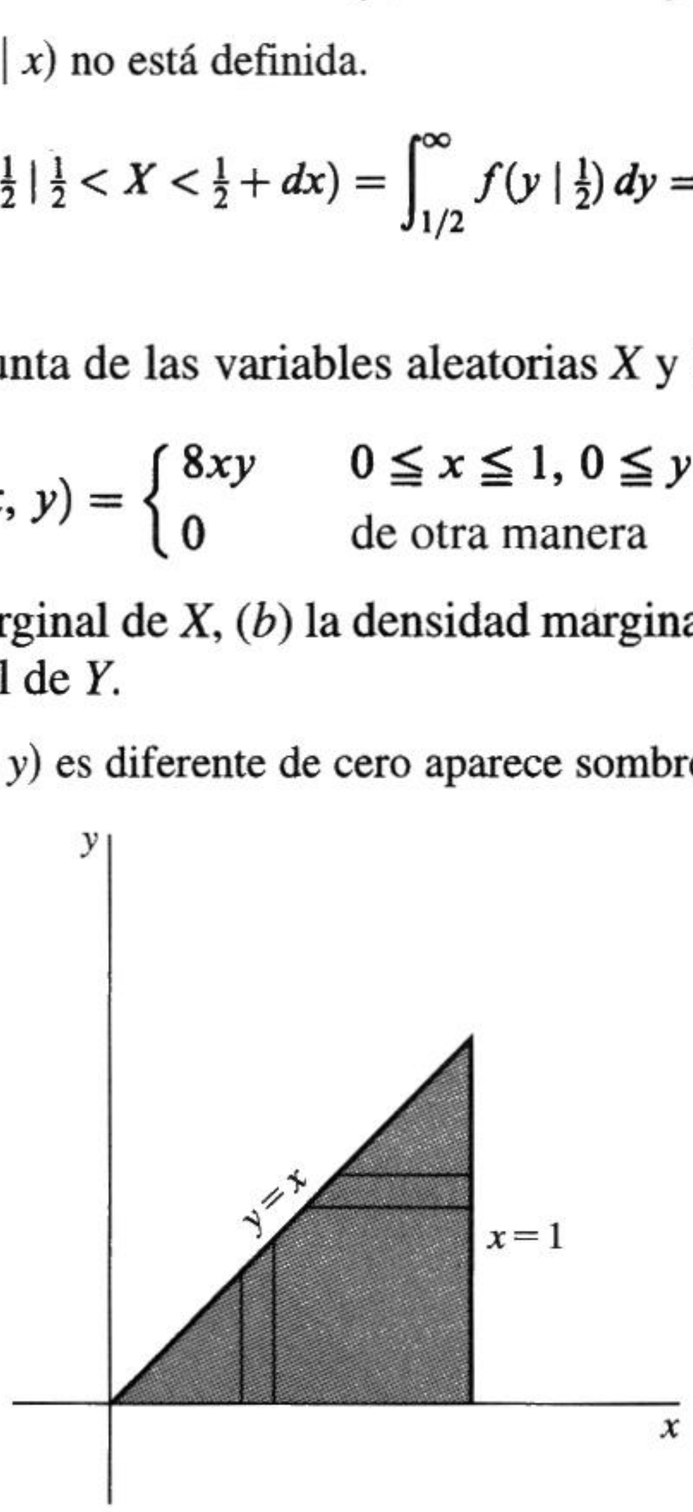


Fig. 2-17

(*a*) Para obtener la densidad marginal de X, fijamos x y la integramos con respecto a y desde 0 hasta x, como se indica en la franja vertical de la figura 2-17. El resultado es

$$f_1(x) = \int_{y=0}^{x} 8xy\,dy = 4x^3$$

para $0 < x < 1$. Para los demás valores de x, $f_1(x) = 0$.

(*b*) De igual manera, la densidad marginal de Y se obtiene fijando y e integrándola con respecto a x desde $x = y$ hasta $x = 1$, como se indica por la franja horizontal en la figura 2-17. El resultado es, para $0 < y < 1$,

$$f_2(y) = \int_{x=y}^{1} 8xy\,dx = 4y(1 - y^2)$$

Para los demás valores de y, $f_2(y) = 0$.

(*c*) La función de densidad condicional de X es, para $0 < y < 1$,

$$f_1(x \mid y) = \frac{f(x, y)}{f_2(y)} = \begin{cases} 2x/(1 - y^2) & y \leqq x \leqq 1 \\ 0 & \text{otro } x \end{cases}$$

La función de densidad condicional no está definida cuando $f_2(y) = 0$.

(*d*) La función de densidad condicional de Y es, para $0 < x < 1$,

$$f_2(y \mid x) = \frac{f(x, y)}{f_1(x)} = \begin{cases} 2y/x^2 & 0 \leqq y \leqq x \\ 0 & \text{otro } y \end{cases}$$

La función de densidad condicional no está definida cuando $f_1(x) = 0$.

Verificación: $\int_0^1 f_1(x)\,dx = \int_0^1 4x^3\,dx = 1$, $\int_0^1 f_2(y)\,dy = \int_0^1 4y(1 - y^2)\,dy = 1$

$$\int_y^1 f_1(x \mid y)\,dx = \int_y^1 \frac{2x}{1 - y^2}\,dx = 1$$

$$\int_0^x f_2(y \mid x)\,dy = \int_0^x \frac{2y}{x^2}\,dy = 1$$

2.30. Determine si las variables aleatorias del problema 2.29 son independientes.

En la región sombreada de la figura 2-17, $f(x, y) = 8xy$, $f_1(x) = 4x^3$, $f_2(y) = 4y(1 - y^2)$. De ahí que $f(x, y) \neq f_1(x)f_2(y)$, y X y Y son independientes.

Se debe observar que a partir de $f(x, y) = 8xy$ no puede deducirse que $f(x, y)$ se pueda expresar como una función de x sola por una función de y sola. Esto se debe a la restricción $0 \leqq y \leqq x$. Si ésta se remplazara por alguna restricción de y que no dependiera de x (como en el problema 2.21), tal conclusión sería válida.

APLICACIONES CON PROBABILIDAD GEOMÉTRICA

2.31. Jugando dardos una persona encuentra que la probabilidad de que un dardo caiga entre r y $r + dr$ es

$$P(r \leqq R \leqq r + dr) = c\left[1 - \left(\frac{r}{a}\right)^2\right] dr$$

Aquí, R es la distancia del impacto desde el centro del blanco o diana, c es una constante y a es el radio del blanco (*véase* figura 2-18). Encuentre la probabilidad de acertar en la diana, la cual suponemos que tiene un radio b. Suponga que siempre da en el blanco.

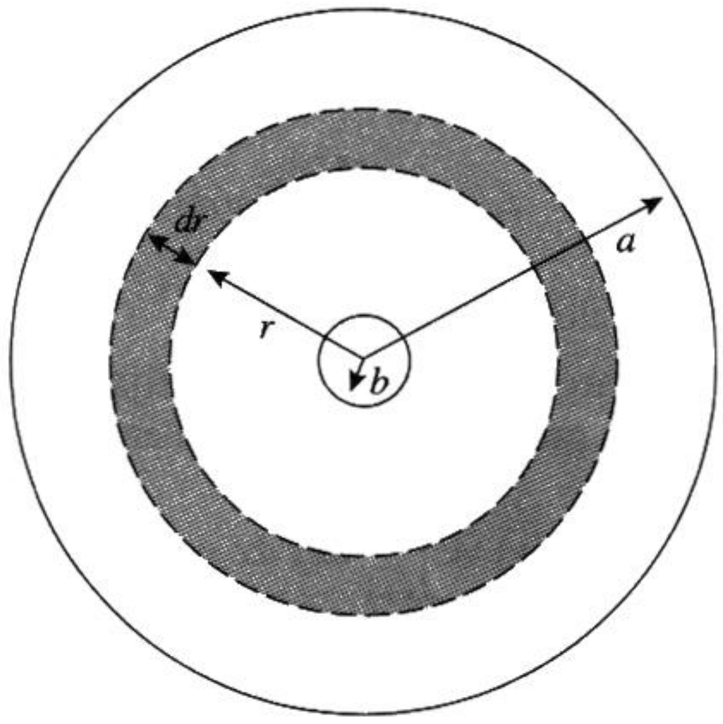


Fig. 2-18

La función de densidad está dada por

$$f(r) = c\left[1 - \left(\frac{r}{a}\right)^2\right]$$

Puesto que siempre da en el blanco, tenemos

$$c\int_0^a \left[1 - \left(\frac{r}{a}\right)^2\right] dr = 1$$

a partir de lo cual tenemos $c = 3/2a$. Entonces, la probabilidad de dar en la diana es

$$\int_0^b f(r)\, dr = \frac{3}{2a}\int_0^b \left[1 - \left(\frac{r}{a}\right)^2\right] dr = \frac{b(3a^2 - b^2)}{2a^3}$$

2.32. Se seleccionan al azar dos puntos en el intervalo $0 \leqq x \leqq 1$. Determine la probabilidad de que la suma de sus cuadrados sea menor que 1.

Sean X y Y variables aleatorias asociadas con los puntos dados. Puesto que supusimos que intervalos iguales tienen probabilidades iguales, las funciones de densidad de X y Y están dadas respectivamente por

$$(1) \qquad f_1(x) = \begin{cases} 1 & 0 \leqq x \leqq 1 \\ 0 & \text{de otra manera} \end{cases} \qquad f_2(y) = \begin{cases} 1 & 0 \leqq y \leqq 1 \\ 0 & \text{de otra manera} \end{cases}$$

Entonces, como X y Y son independientes, la función de densidad conjunta está dada por

$$(2) \qquad f(x, y) = f_1(x)f_2(y) = \begin{cases} 1 & 0 \leqq x \leqq 1,\ 0 \leqq y \leqq 1 \\ 0 & \text{de otra manera} \end{cases}$$

Se deduce que la probabilidad requerida está dada por

$$(3) \qquad P(X^2 + Y^2 \leqq 1) = \iint_{\mathscr{R}} dx\, dy$$

donde $\mathscr{R}$ es la región definida por $x^2 + y^2 \leqq 1$, $x \geqq 0$, $y \geqq 0$, lo cual equivale a un cuarto de un círculo con radio 1 (figura 2-19). Ahora, dado que (*3*) representa el área de $\mathscr{R}$, vemos que la probabilidad requerida es $\pi/4$.

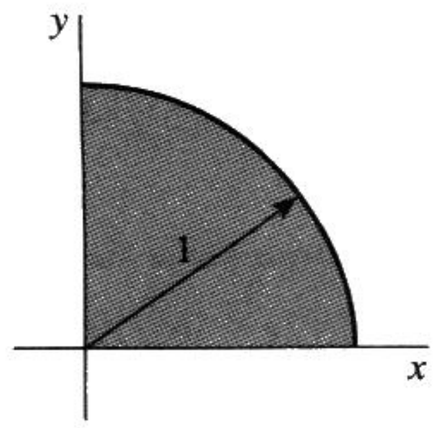


Fig. 2-19

PROBLEMAS DIVERSOS

2.33. Supongamos que las variables aleatorias X y Y tienen una función de densidad conjunta dada por

$$f(x, y) = \begin{cases} c(2x + y) & 2 < x < 6,\ 0 < y < 5 \\ 0 & \text{de otra manera} \end{cases}$$

Encuentre (a) la constante c, (b) las funciones de distribución marginal para X y Y, (c) las funciones de densidad marginal para X y Y, (d) $P(3 < X < 4,\ Y > 2)$, (e) $P(X > 3)$, (f) $P(X + Y > 4)$, (g) la función de distribución conjunta, (h) si X y Y son independientes.

(a) La probabilidad total está dada por

$$\int_{x=2}^{6} \int_{y=0}^{5} c(2x + y)\, dx\, dy = \int_{x=2}^{6} c\left(2xy + \frac{y^2}{2}\right)\Bigg|_0^5 dx$$

$$= \int_{x=2}^{6} c\left(10x + \frac{25}{2}\right) dx = 210c$$

para igualarla a 1, debemos tener $c = 1/210$.

(b) La función de distribución marginal para X es

$$F_1(x) = P(X \leqq x) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{\infty} f(u, v)\, du\, dv$$

$$= \begin{cases} \displaystyle\int_{u=-\infty}^{x} \int_{v=-\infty}^{\infty} 0\, du\, dv = 0 & x < 2 \\ \displaystyle\int_{u=2}^{x} \int_{v=0}^{5} \frac{2u + v}{210}\, du\, dv = \frac{2x^2 + 5x - 18}{84} & 2 \leqq x < 6 \\ \displaystyle\int_{u=2}^{6} \int_{v=0}^{5} \frac{2u + v}{210}\, du\, dv = 1 & x \geqq 6 \end{cases}$$

La función de distribución marginal para Y es

$$F_2(y) = P(Y \leqq y) = \int_{u=-\infty}^{\infty} \int_{v=-\infty}^{y} f(u, v)\, du\, dv$$

$$= \begin{cases} \displaystyle\int_{u=-\infty}^{\infty} \int_{v=-\infty}^{y} 0\, du\, dv = 0 & y < 0 \\ \displaystyle\int_{u=2}^{6} \int_{v=0}^{y} \frac{2u + v}{210}\, du\, dv = \frac{y^2 + 16y}{105} & 0 \leqq y < 5 \\ \displaystyle\int_{u=2}^{6} \int_{v=0}^{5} \frac{2u + v}{210}\, du\, dv = 1 & y \geqq 5 \end{cases}$$

(*c*) La función de densidad marginal para X es, a partir de la parte (*b*),

$$f_1(x) = \frac{d}{dx} F_1(x) = \begin{cases} (4x+5)/84 & 2 < x < 6 \\ 0 & \text{de otra manera} \end{cases}$$

La función de densidad marginal para Y es, a partir de la parte (*b*),

$$f_2(y) = \frac{d}{dy} F_2(y) = \begin{cases} (2y+16)/105 & 0 < y < 5 \\ 0 & \text{de otra manera} \end{cases}$$

(*d*)
$$P(3 < X < 4,\ Y > 2) = \frac{1}{210} \int_{x=3}^{4} \int_{y=2}^{5} (2x+y)\, dx\, dy = \frac{3}{20}$$

(*e*)
$$P(X > 3) = \frac{1}{210} \int_{x=3}^{6} \int_{y=0}^{5} (2x+y)\, dx\, dy = \frac{23}{28}$$

(*f*)
$$P(X + Y > 4) = \iint_{\mathscr{R}} f(x,y)\, dx\, dy$$

donde $\mathscr{R}$ es la región sombreada oscura de la figura 2-20. Aunque esto se puede resolver, es más fácil usar el hecho de que

$$P(X + Y > 4) = 1 - P(X + Y \leqq 4) = 1 - \iint_{\mathscr{R}'} f(x,\, y)\, dx\, dy$$

donde $\mathscr{R}'$ es la región sombreada clara de la figura 2-20. Tenemos

$$P(X + Y \leqq 4) = \frac{1}{210} \int_{x=2}^{4} \int_{y=0}^{4-x} (2x+y)\, dx\, dy = \frac{2}{35}$$

Por lo tanto $P(X + Y) = 33/35$

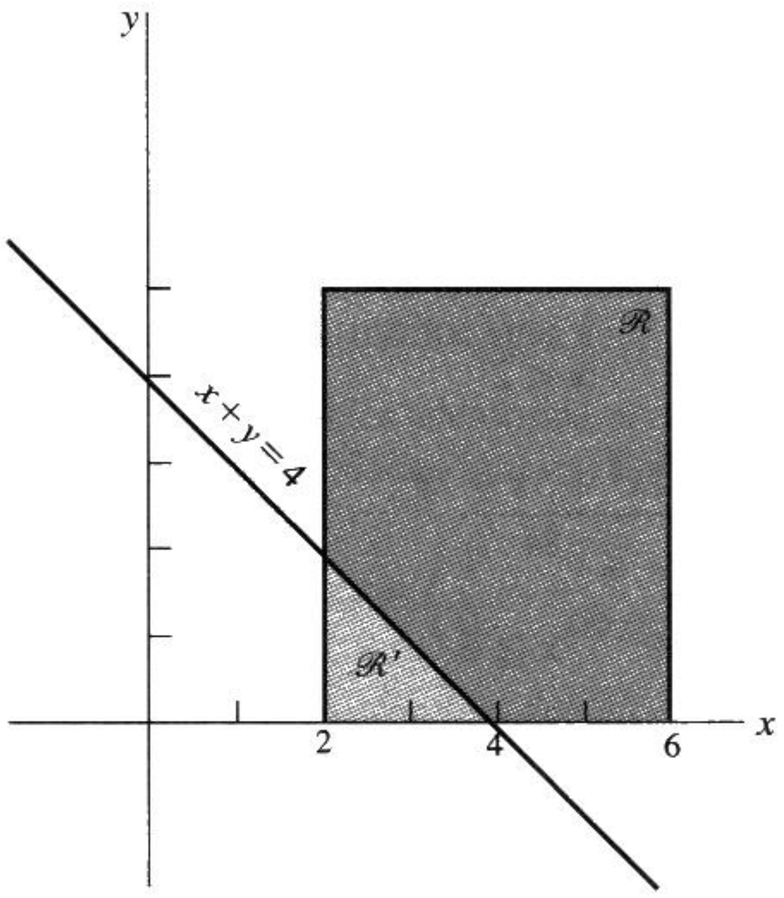


Fig. 2-20

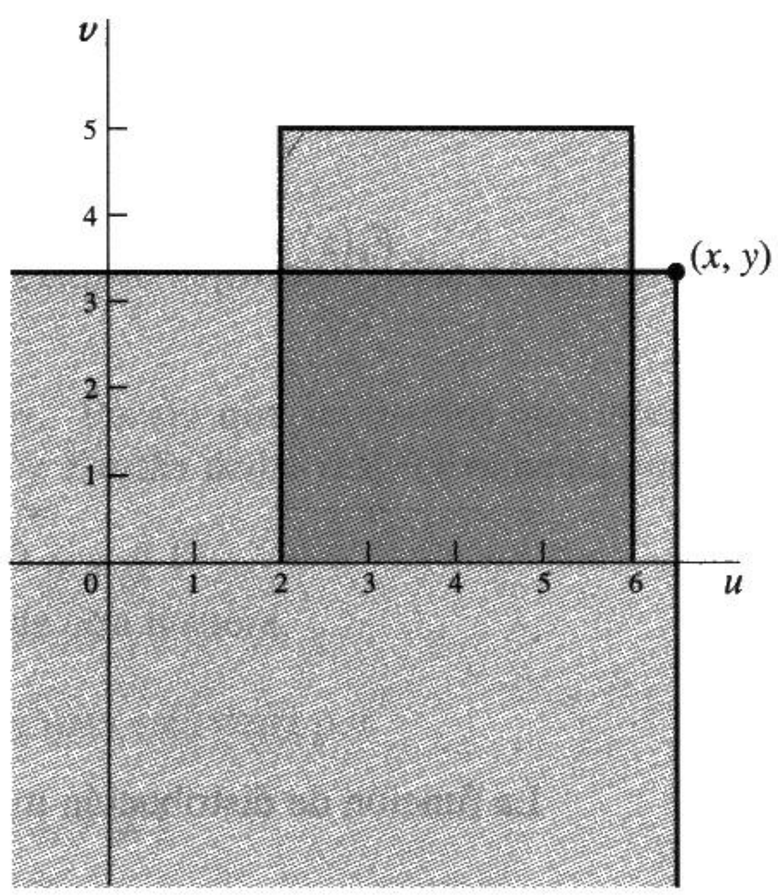


Fig. 2-21

(*g*) La función de distribución conjunta es

$$F(x,y) = P(X \leqq x,\, Y \leqq y) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{y} f(u,v)\, du\, dv$$

En el plano uv (figura 2-21) la región de integración es la intersección del plano del cuadrante $u \leqq x$, $v \leqq y$, y el rectángulo $2 < y < 6$, $0 < v < 5$ [en la cual $f(u, v)$ no es cero]. Para (x, y) ubicadas como aparecen en la figura, tenemos

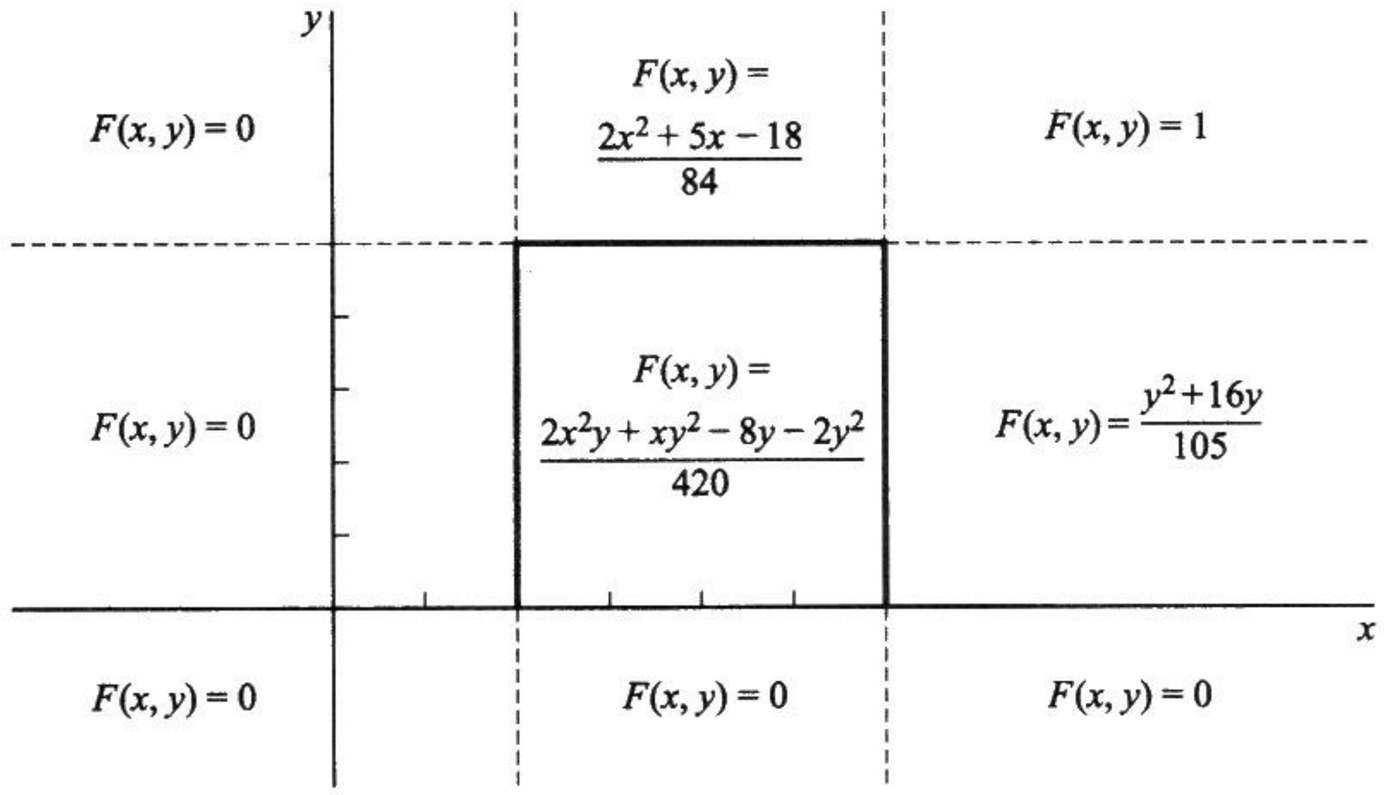


Fig. 2-22

$$F(x, y) = \int_{u=2}^{6} \int_{v=0}^{y} \frac{2u + v}{210}\, du\, dv = \frac{16y + y^2}{105}$$

Donde (x, y) se ubica dentro del rectángulo, obtenemos otra expresión, etc. Los resultados completos se muestran en la figura 2-22.

(h) Las variables aleatorias son dependientes, ya que

$$f(x, y) \neq f_1(x) f_2(y)$$

o de manera equivalente, $F(x, y) \neq F_1(x) F_2(y)$.

2.34. Sea X la función de densidad

$$f(x) = \begin{cases} 6x(1 - x) & 0 < x < 1 \\ 0 & \text{de otra manera} \end{cases}$$

Encuentre una función $Y = h(X)$, que tenga una función de densidad

$$g(y) = \begin{cases} 12y^3(1 - y^2) & 0 < y < 1 \\ 0 & \text{de otra manera} \end{cases}$$

Suponemos que la función desconocida h es tal que los intervalos $X \leqq x$ y $Y \leqq y + h(x)$ corresponden a una relación continua uno a uno. Entonces, $P(X \leqq x) = P(Y \leqq y)$, es decir, las funciones de distribución de X y Y deben ser iguales. Así, para $0 < x, y < 1$,

$$\int_0^x 6u(1 - u)\, du = \int_0^y 12v^3(1 - v^2)\, dv$$

o

$$3x^2 - 2x^3 = 3y^4 - 2y^6$$

Por inspección, $x = y^2$ o $y = h(x) = +\sqrt{x}$ es una solución, y esta solución tiene las propiedades deseadas. Así, $Y = +\sqrt{X}$.

2.35. Encuentre la función de densidad de $U = XY$ si la función de densidad conjunta de X y Y es $f(x, y)$.

Método 1.

Sea $U = XY$ y $V = X$, a las cuales les corresponde $u = xy$, $v = x$ o $x = v$, $y = u/v$. Entonces el jacobiano está dado por

$$J = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ v^{-1} & -uv^{-2} \end{vmatrix} = -v^{-1}$$

Por lo tanto la función de densidad conjunta de U y V es

$$g(u, v) = \frac{1}{|v|} f\left(v, \frac{u}{v}\right)$$

a partir de la cual la función de densidad marginal de U se obtiene como

$$g(u) = \int_{-\infty}^{\infty} g(u, v)\, dv = \int_{-\infty}^{\infty} \frac{1}{|v|} f\left(v, \frac{u}{v}\right) dv$$

Método 2.

La función de distribución de U es

$$G(u) = \iint_{xy \leqq u} f(x, y)\, dx\, dy$$

Para $u \geqq 0$, la región de integración aparece sombreada en la figura 2-23. Vemos que

$$G(u) = \int_{-\infty}^{0} \left[\int_{u/x}^{\infty} f(x, y)\, dy \right] dx + \int_{0}^{\infty} \left[\int_{-\infty}^{u/x} f(x, y)\, dy \right] dx$$

Derivando con respecto a u, obtenemos

$$g(u) = \int_{-\infty}^{0} \left(\frac{-1}{x}\right) f\left(x, \frac{u}{x}\right) dx + \int_{0}^{\infty} \frac{1}{x} f\left(x, \frac{u}{x}\right) dx = \int_{-\infty}^{\infty} \frac{1}{|x|} f\left(x, \frac{u}{x}\right) dx$$

El mismo resultado se obtiene para $u < 0$, cuando la región de integración está delimitada por la hipérbola punteada de la figura 2-23.

2.36. Un piso tiene líneas paralelas a distancias iguales l entre sí. Se lanza al azar una aguja de longitud $a < l$ al suelo. Encuentre la probabilidad de que la aguja interseque una línea. (Este problema se conoce como el *problema de la aguja de Buffon*.)

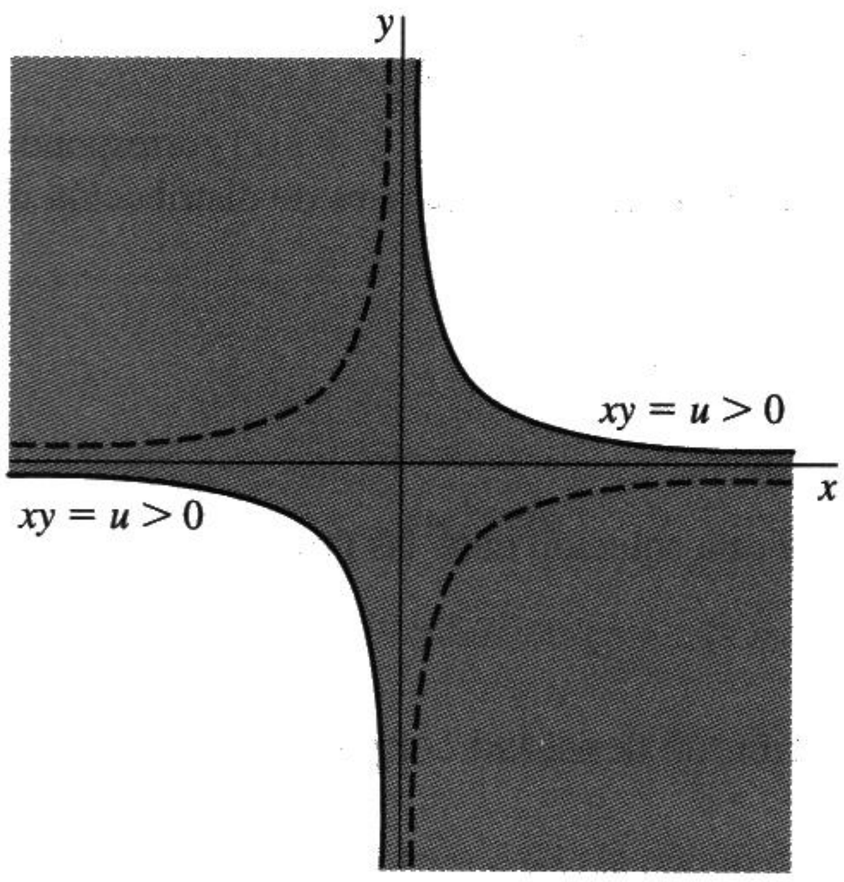


Fig. 2-23

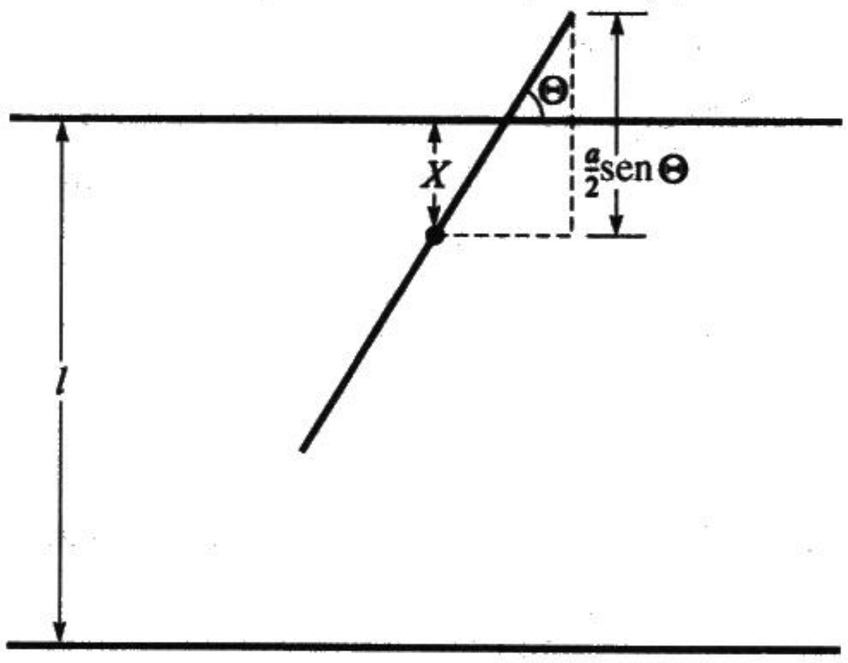


Fig. 2-24

Sea X una variable aleatoria que da la distancia desde el punto medio de la aguja a la línea más cercana (figura 2-24). Sea Θ una variable aleatoria que da el ángulo agudo entre la aguja (o su extensión) y la línea. Denominamos x y θ cualquier valor particular de X y Θ. Se observa que X puede tomar cualquier valor entre 0 y $l/2$, de manera que $0 \leqq x \leqq l/2$. Θ también puede tomar cualquier valor entre 0 y $\pi/2$. Se deduce que

$$P(x < X \leqq x + dx) = \frac{2}{l}\,dx \qquad P(\theta \leqq \Theta + d\theta) = \frac{2}{\pi}\,d\theta$$

es decir, las funciones de densidad de X y Θ están dadas por $f_1(x) = 2/l$, $f_2(\theta) = 2/\pi$. Como verificación, observamos que

$$\int_0^{l/2} \frac{2}{l}\,dx = 1 \quad ; \quad \int_0^{\pi/2} \frac{2}{\pi}\,d\theta = 1$$

Dado que X y Θ son independientes, la función de densidad conjunta es

$$f(x, \theta) = \frac{2}{l} \cdot \frac{2}{\pi} = \frac{4}{l\pi}$$

A partir de la figura 2-24 se ve que en realidad la aguja toca la línea cuando $X \leqq (a/2)\ \text{sen}\ \Theta$. La probabilidad de este evento está dada por

$$\frac{4}{l\pi} \int_{\theta=0}^{\pi/2} \int_{x=0}^{(a/2)\,\text{sen}\,\theta} dx\,d\theta = \frac{2a}{l\pi}$$

Cuando la expresión anterior se iguala a la frecuencia de aciertos observados en experimentos reales, se obtienen valores precisos de π. Esto indica que el modelo probabilístico descrito anteriormente es adecuado.

2.37. Dos personas acuerdan encontrarse entre las 2:00 p. m. y las 3:00 p. m., con el compromiso de que cada una esperará a la otra no más de 15 minutos. ¿Cuál es la probabilidad de que estas personas se encuentren?

Sean X y Y variables aleatorias que representan tiempos de llegada, medidos en fracciones de una hora después de las 2:00 p. m., de ambas personas. Suponiendo que intervalos de tiempo iguales tienen iguales probabilidades de llegada, las funciones de densidad de X y Y están dadas respectivamente por

$$f_1(x) = \begin{cases} 1 & 0 \leqq x \leqq 1 \\ 0 & \text{de otra manera} \end{cases}$$

$$f_2(y) = \begin{cases} 1 & 0 \leqq y \leqq 1 \\ 0 & \text{de otra manera} \end{cases}$$

Entonces, dado que X y Y son independientes, la función de densidad conjunta es

$$(1) \qquad f(x, y) = f_1(x) f_2(y) = \begin{cases} 1 & 0 \leqq x \leqq 1,\ 0 \leqq y \leqq 1 \\ 0 & \text{de otra manera} \end{cases}$$

Dado que 15 minutos = 1/4 hora, la probabilidad requerida es

$$(2) \qquad P\left(|X - Y| \leqq \frac{1}{4}\right) = \iint_{\mathscr{R}} dx\,dy$$

donde $\mathscr{R}$ es la región sombreada de la figura 2-25. El lado derecho de (2) es el área de esta región, la cual es igual a $1 - (\frac{3}{4})(\frac{3}{4}) = \frac{7}{16}$, dado que el cuadrado tiene un área de 1, mientras que los dos triángulos de las esquinas tienen áreas de $(\frac{1}{2})(\frac{3}{4})(\frac{3}{4})$ cada uno. Así, la probabilidad requerida es 7/16.

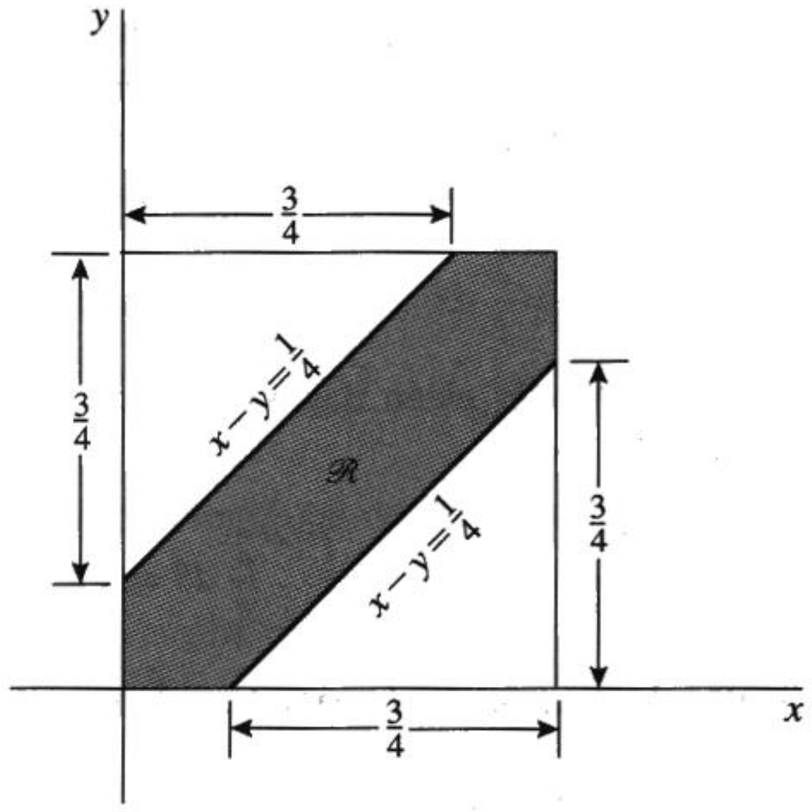


Fig. 2-25

Problemas suplementarios

VARIABLES ALEATORIAS DISCRETAS Y DISTRIBUCIONES DE PROBABILIDAD

2.38. Una moneda se lanza tres veces. Si X es una variable aleatoria que indica el número de caras que aparecen, elabore una tabla que muestre la distribución de probabilidad de X.

2.39. Una urna contiene 5 canicas blancas y 3 negras. Si se sacan 2 canicas al azar sin remplazo y X indica el número de canicas blancas, encuentre la probabilidad de X.

2.40. Desarrolle el problema 2.39 si las canicas se sacan con remplazo.

2.41. Sea Z la variable aleatoria que indica el número de caras menos el número de sellos en dos lanzamientos de una moneda balanceada. Encuentre la distribución de probabilidad de Z. Compare el resultado con los ejemplos 2.1 y 2.2.

2.42. Sea X una variable aleatoria que indica el número de ases al retirar 4 cartas al azar de una baraja común de 52 cartas. Elabore una tabla que muestre la distribución de probabilidad de X.

FUNCIONES DE DISTRIBUCIÓN DISCRETA

2.43. En la tabla 2-7 se muestra la función de probabilidad de una variable aleatoria X. Elabore una tabla que muestre la función de distribución de X.

Tabla 2-7

x	1	2	3
$f(x)$	1/2	1/3	1/6

Tabla 2-8

x	1	2	3	4
$F(x)$	1/8	3/8	3/4	1

2.44. Obtenga la función de distribución para (*a*) el problema 2.38, (*b*) el problema 2.39, (*c*) el problema 2.40.

2.45. Obtenga la función de distribución para (*a*) el problema 2.41, (*b*) el problema 2.42.

2.46. La tabla 2-8 muestra la función de distribución de una variable aleatoria X. Determine (a) la función de probabilidad, (b) $P(1 \leqq X \leqq 3)$, (c) $P(X \geqq 2)$, (d) $P(X < 3)$, (e) $P(X > 1.4)$.

VARIABLES ALEATORIAS CONTINUAS Y DISTRIBUCIONES DE PROBABILIDAD

2.47. Una variable aleatoria X tiene función de densidad

$$f(x) = \begin{cases} ce^{-3x} & x > 0 \\ 0 & x \leqq 0 \end{cases}$$

Encuentre (a) la constante c, (b) $P(1 < X < 2)$, (c) $P(X \geqq 3)$, (d) $P(X < 1)$.

2.48. Encuentre la función de distribución para la variable aleatoria del problema 2.47. Represente gráficamente las funciones de distribución y densidad de manera que describa la relación entre ellas.

2.49. Una variable aleatoria X tiene función de densidad

$$f(x) = \begin{cases} cx^2 & 1 \leqq x \leqq 2 \\ cx & 2 < x < 3 \\ 0 & \text{de otra manera} \end{cases}$$

Encuentre (a) la constante c, (b) $P(X > 2)$, (c) $P(1/2 < X < 3/2)$

2.50. Encuentre la función de distribución para la variable aleatoria X del problema 2.49.

2.51. La función de distribución de una variable aleatoria X está dada por

$$F(x) = \begin{cases} cx^3 & 0 \leqq x < 3 \\ 1 & x \geqq 3 \\ 0 & x < 0 \end{cases}$$

Si $P(X = 3) = 0$, encuentre (a) la constante c, (b) la función de densidad, (c) $P(X > 1)$, (d) $P(1 < X < 2)$.

2.52. ¿Puede ser la siguiente función

$$F(x) = \begin{cases} c(1 - x^2) & 0 \leqq x \leqq 1 \\ 0 & \text{de otra manera} \end{cases}$$

una función de distribución? Explique.

2.53. Sea X una variable aleatoria con función de densidad

$$f(x) = \begin{cases} cx & 0 \leqq x \leqq 2 \\ 0 & \text{de otra manera} \end{cases}$$

Encuentre (a) el valor de la constante c, (b) $P(\frac{1}{2} < X < \frac{3}{2})$, (c) $P(X > 1)$, (d) la función de distribución.

DISTRIBUCIONES CONJUNTAS Y VARIABLES INDEPENDIENTES

2.54. La función de probabilidad conjunta de dos variables aleatorias discretas X y Y está dada por $f(x, y) = cxy$ para $x = 1, 2, 3$ y $y = 1, 2, 3$, y de otra manera igual a cero. Encuentre (a) la constante c, (b) $P(X = 2, Y = 3)$, (c) $P(1 \leqq 2, Y \leqq 2)$, (d) $P(X \geqq 2)$, (e) $P(Y < 2)$, (f) $P(X = 1)$, (g) $P(Y = 3)$.

2.55. Encuentre las funciones de probabilidad marginal de (*a*) X, y (*b*) Y, para las variables aleatorias del problema 2.54. (*c*) Determine si X y Y son independientes.

2.56. Sean X y Y variables aleatorias continuas con función de densidad.

$$f(x, y) = \begin{cases} c(x^2 + y^2) & 0 \leqq x \leqq 1,\ 0 \leqq y \leqq 1 \\ 0 & \text{en otro caso} \end{cases}$$

Determine (*a*) la constante c, (*b*) $P(X < \frac{1}{2},\ Y > \frac{1}{2})$, (*c*) $P(\frac{1}{4} < X < \frac{3}{4})$, (*d*) $P(Y < \frac{1}{2})$, (*e*) si X y Y son independientes.

2.57. Encuentre las funciones de distribución marginal (*a*) de X, y (*b*) de Y, para la función de densidad del problema 2.56.

DISTRIBUCIONES CONDICIONALES Y FUNCIONES DE DENSIDAD

2.58. Encuentre la función de densidad condicional (*a*) de X dado Y, (*b*) de Y dado X, para la distribución del problema 2.54.

2.59. Sea

$$f(x, y) = \begin{cases} x + y & 0 \leqq x \leqq 1,\ 0 \leqq y \leqq 1 \\ 0 & \text{en otro caso} \end{cases}$$

Encuentre la función de densidad condicional de (*a*) X dado Y, (*b*) Y dado X.

2.60. Encuentre la densidad condicional de (*a*) X dado Y, (*b*) Y dado X, para la distribución del problema 2.56.

2.61. Sea

$$f(x, y) = \begin{cases} e^{-(x+y)} & x \geqq 0,\ y \geqq 0 \\ 0 & \text{de otra manera} \end{cases}$$

la función de densidad conjunta de X y Y. Encuentre la función de densidad condicional de (*a*) X dado Y, (*b*) Y dado X.

CAMBIO DE VARIABLES

2.62. Si X tiene la función de densidad

$$f(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & x \leqq 0 \end{cases}$$

Encuentre la función de densidad de $Y = X^2$.

2.63. (*a*) Si la función de densidad de X es $f(x)$, encuentre la función de densidad de X^3. (*b*) Ilustre el resultado de la parte (*a*) seleccionando

$$f(x) = \begin{cases} 2e^{-2x} & x \geqq 0 \\ 0 & x < 0 \end{cases}$$

y verifique la respuesta.

2.64. Si X tiene función de densidad $f(x) = (2\pi)^{-1/2} e^{-x^2/2}$, $-\infty < x < \infty$, encuentre la función de densidad de $Y = X^2$.

2.65. Verifique que la integral de $g_1(u)$ en el método 1 del problema 2.21 es igual a 1.

2.66. Si la densidad de X es $f(x) = 1/\pi(x^2 + 1)$, $-\infty < x < \infty$, encuentre la densidad de $Y = \tan^{-1} X$.

2.67. Complete el trabajo necesario para encontrar $g_1(u)$ en el método 2 del problema 2.21 y verifique su respuesta.

2.68. Si la densidad de X es

$$f(x) = \begin{cases} 1/2 & -1 < x < 1 \\ 0 & \text{de otra manera} \end{cases}$$

Encuentre la densidad de (*a*) $3X - 2$, (*b*) $X^3 + 1$.

2.69. Verifique por integración directa la función de densidad conjunta encontrada en el problema 2.22.

2.70. Si X y Y tienen función de densidad conjunta

$$f(x, y) = \begin{cases} e^{-(x+y)} & x \geqq 0, y \geqq 0 \\ 0 & \text{de otra manera} \end{cases}$$

Si $U = X/Y$, $V = X + Y$, encuentre la función de densidad conjunta de U y V.

2.71. Use el problema 2.22 para encontrar la función de densidad de (*a*) $U = XY^2$, (*b*) $V = X^2Y$.

2.72. Sean X y Y variables aleatorias con función de densidad conjunta $f(x, y) = (2\pi)^{-1}e^{-(x^2 + y^2)}$, $-\infty < x < \infty$, $-\infty < y < \infty$. Si R y Θ son variables aleatorias nuevas tal que $X = R\cos\Theta$, $Y = R\text{sen}\Theta$, demuestre que la función de R es

$$g(r) = \begin{cases} re^{-r^2/2} & r \geqq 0 \\ 0 & r < 0 \end{cases}$$

2.73. Sea

$$f(x, y) = \begin{cases} 1 & 0 \leqq x \leqq 1, 0 \leqq y \leqq 1 \\ 0 & \text{de otra manera} \end{cases}$$

la función de densidad conjunta de X y Y. Encuentre la función de densidad de $Z = XY$.

CONVOLUCIONES

2.74. Sean X y Y variables aleatorias independientes idénticamente distribuidas con función de densidad

$$f(t) = \begin{cases} 1 & 0 \leqq t \leqq 1 \\ 0 & \text{de otra manera} \end{cases}$$

Encuentre la función de densidad de $X + Y$ y verifique su respuesta.

2.75. Sean X y Y variables aleatorias independientes idénticamente distribuidas con función de densidad

$$f(t) = \begin{cases} e^{-t} & t \geqq 0 \\ 0 & \text{de otra manera} \end{cases}$$

Encuentre la función de densidad de $X + Y$ y verifique su respuesta.

2.76. Desarrolle el problema 2.21 haciendo primero la transformación $2Y = Z$, y luego usando convoluciones para encontrar la función de densidad de $U = X + Z$.

2.77. Si las variables aleatorias independientes X_1 y X_2 están idénticamente distribuidas con función de densidad

$$f(t) = \begin{cases} te^{-t} & t \geqq 0 \\ 0 & t < 0 \end{cases}$$

encuentre la función de densidad de $X_1 + X_2$.

APLICACIONES CON PROBABILIDAD GEOMÉTRICA

2.78. Se escogen al azar dos puntos de un segmento de recta cuya longitud es $a > 0$. Encuentre la probabilidad de que tres segmentos de recta con características iguales sean los lados de un triángulo.

2.79. Se sabe que un autobús llegará al azar a cierto lugar en cualquier momento entre las 3 p. m. y las 3:30 p. m. Un hombre decide que irá al azar a dicho lugar en ese lapso y esperará el autobús máximo 5 minutos. Si él pierde el autobús, tomará un taxi. ¿Cuál es la probabilidad de que tome un taxi?

2.80. Dos segmentos de recta, AB y CD, tienen longitudes de 8 y 6 unidades, respectivamente. Se van a escoger al azar dos puntos P y Q de los segmentos AB y CD, respectivamente. Demuestre que la probabilidad de que el área de un triángulo tendrá una altura AP y de que la base CQ será mayor que 12 unidades cuadradas, es igual a $(1 - \ln 2)/2$.

PROBLEMAS DIVERSOS

2.81. Suponga que que $f(x) = c/3^x$, $x = 1, 2, \ldots$, es la función de probabilidad para una variable aleatoria X. (a) Determine c. (b) Encuentre la función de distribución. (c) Elabore una gráfica de la función de probabilidad y de la función de distribución. (d) Encuentre $P(2 \leqq X < 5)$. (e) Encuentre $P(X \geqq 3)$.

2.82. Suponga que

$$f(x) = \begin{cases} cxe^{-2x} & x \geqq 0 \\ 0 & \text{de otra manera} \end{cases}$$

es la función de densidad para una variable aleatoria X. (a) Determine c. (b) Encuentre la función de distribución. (c) Elabore una gráfica con la función de densidad y la función de distribución. (d) Encuentre $P(X \geqq 1)$. (e) Encuentre $P(2 \leqq X \leqq 3)$.

2.83. La función de probabilidad de una variable aleatoria X está dada por

$$f(x) = \begin{cases} 2p & x = 1 \\ p & x = 2 \\ 4p & x = 3 \\ 0 & \text{de otra manera} \end{cases}$$

donde p es una constante. Encuentre (a) $P(0 \leqq X < 3)$, (b) $P(X > 1)$.

2.84. (a) Demuestre que para una constante apropiada c,

$$F(x) = \begin{cases} 0 & x \leqq 0 \\ c(1 - e^{-x})^2 & x > 0 \end{cases}$$

es la función de distribución para una variable aleatoria X, y encuentre c. (b) Encuentre $P(1 < X < 2)$.

2.85. Una variable aleatoria X tiene función de densidad

$$f(x) = \begin{cases} \frac{3}{2}(1 - x^2) & 0 \leqq x \leqq 1 \\ 0 & \text{de otra manera} \end{cases}$$

Encuentre la función de densidad de la variable aleatoria $Y = X^2$ y verifique su respuesta.

2.86. Dos variables aleatorias independientes, X y Y, tienen las siguientes funciones de densidad, respectivamente

$$f(x) = \begin{cases} c_1 e^{-2x} & x > 0 \\ 0 & x \leqq 0 \end{cases} \qquad g(y) = \begin{cases} c_2 y e^{-3y} & y > 0 \\ 0 & y \leqq 0 \end{cases}$$

Encuentre (a) c_1 y c_2, (b) $P(X + Y > 1)$, (c) $P(1 < X < 2, Y \geqq 1)$, (d) $P(1 < X < 2)$, (e) $P(Y \geqq 1)$.

2.87. ¿Qué relación existe entre las respuestas a (c), (d) y (e) delproblema 2.86? Justifique su respuesta.

2.88. Sean X y Y variables aleatorias que tienen función de densidad conjunta

$$f(x, y) = \begin{cases} c(2x + y) & 0 < x < 1, 0 < y < 2 \\ 0 & \text{de otra manera} \end{cases}$$

Encuentre (a) la constante c, (b) $P(X > \frac{1}{2}, \ Y < \frac{3}{2})$, (c) la función de densidad (marginal) de X, (d) la función de densidad (marginal) de Y.

2.89. En el problema 2.88, ¿es $P(X > \frac{1}{2}, \ Y < \frac{3}{2}) = P(X > \frac{1}{2})P(Y < \frac{3}{2})$? ¿Por qué?

2.90. Encuentre la función de densidad del problema 2.86 (a) de X^2, (b) de $X + Y$.

2.91. Si X y Y tienen función de densidad conjunta

$$f(x, y) = \begin{cases} 1/y & 0 < x < y, 0 < y < 1 \\ 0 & \text{de otra manera} \end{cases}$$

(a) Determine si X y Y son independientes. (b) Encuentre $P(X > \frac{1}{2})$.
(c) Encuentre $P(X < \frac{1}{2}, \ Y > \frac{1}{3})$. (d) Encuentre $P(X + Y > \frac{1}{2})$.

2.92. Generalice (a) el problema 2.74 y (b) el problema 2.75 para tres o más variables.

2.93. Sean X y Y variables aleatorias independientes idénticamente distribuidas con función de densidad $f(u) = (2\pi)^{-1/2} e^{-u^2/2}$, $-\infty < u < \infty$. Encuentre la función de densidad de $Z = X^2 - Y^2$.

2.94. La función de probabilidad conjunta para las variables aleatorias X y Y está dada en la tabla 2-9. (a) Encuentre las funciones de probabilidad marginal de X y Y. (b) Encuentre $P(1 \leqq X < 3, Y \geqq 1)$. (c) Determine si X y Y son independientes.

Tabla 2-9

X \ Y	0	1	2
0	1/18	1/9	1/6
1	1/9	1/18	1/9
2	1/6	1/6	1/18

2.95. Suponga que la función de probabilidad conjunta de las variables aleatorias X y Y está dada por

$$f(x, y) = \begin{cases} cxy & 0 \leqq x \leqq 2, 0 \leqq y \leqq x \\ 0 & \text{de otra manera} \end{cases}$$

(a) Determine si X y Y son independientes. (b) Encuentre $P(\frac{1}{2} < X < 1)$. (c) Encuentre $P(Y \geqq 1)$. (d) Encuentre $P(\frac{1}{2} < X < 1, \ Y \geqq 1)$.

2.96. Sean X y Y variables aleatorias independientes cada una con función de densidad

$$f(u) = \frac{\lambda^u e^{-\lambda}}{u!} \qquad u = 0, 1, 2, \ldots$$

donde $\lambda > 0$. Demuestre que la función de densidad de X + Y es

$$g(u) = \frac{(2\lambda)^u e^{-2\lambda}}{u!} \qquad u = 0, 1, 2, \ldots$$

2.97. Se rompe un palo de longitud L en dos partes. ¿Cuál es la probabilidad de que una parte tenga una longitud de más del doble de la de la otra parte? Indique claramente las suposiciones que debe hacer. Discuta si cree que dichas suposiciones son reales y cómo puede hacer para mejorarlas si no lo son.

2.98. Un piso está hecho de cuadrados de lado l. Se va a arrojar al suelo una aguja de longitud $a < l$. Demuestre que la probabilidad de que la aguja interseque al menos un lado es igual a $a(4l - a)/\pi l^2$.

2.99. Para una aguja de longitud dada, ¿cuál debería ser la longitud del lado del cuadrado del problema 2.98 para que la probabilidad de intersección sea máxima? Explique su respuesta.

2.100. Sea

$$f(x, y, z) = \begin{cases} 24xy^2z^3 & 0 < x < 1,\ 0 < y < 1,\ 0 < z < 1 \\ 0 & \text{de otra manera} \end{cases}$$

la función de densidad conjunta de tres variables aleatorias X, Y y Z. Encuentre (a) $P(X > \frac{1}{2},\ Y < \frac{1}{2},\ Z > \frac{1}{2})$, (b) $P(Z < X + Y)$.

2.101. Una corriente cilíndrica de partículas, de radio a, se dirige hacia un blanco hemisférico ABC, cuyo centro O se indica en la figura 2-26. Suponga que la distribución de las partículas está dada por

$$f(r) = \begin{cases} 1/a & 0 < r < a \\ 0 & \text{de otra manera} \end{cases}$$

donde r es la distancia desde el eje OB. Demuestre que la distribución de las partículas a lo largo del blanco está dada por

$$g(\theta) = \begin{cases} \cos\theta & 0 < \theta < \pi/2 \\ 0 & \text{de otra manera} \end{cases}$$

donde θ es el ángulo que la línea OP (desde O hasta cualquier punto P en el blanco) forma con el eje.

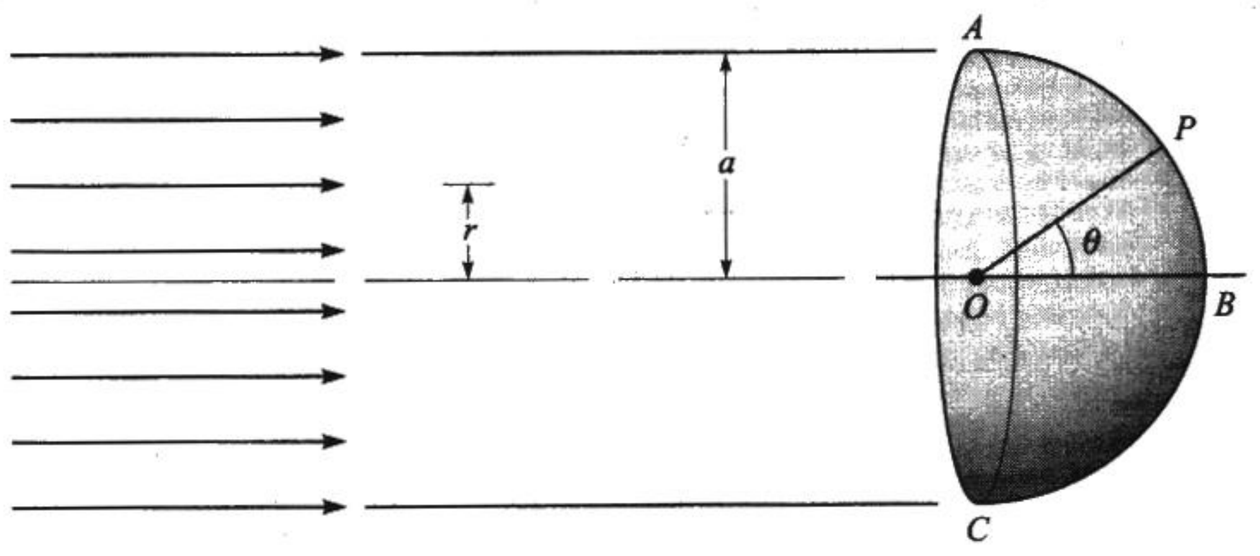


Fig. 2-26

2.102. En el problema 2.101, encuentre la probabilidad de que una partícula dé en el blanco entre $\theta = 0$ y $\theta = \pi/4$.

2.103. Suponga que las variables aleatorias X, Y y Z tienen función de densidad conjunta

$$f(x, y, z) = \begin{cases} 1 - \cos\pi x \cos\pi y \cos\pi z & 0 < x < 1,\ 0 < y < 1,\ 0 < z < 1 \\ 0 & \text{de otra manera} \end{cases}$$

Demuestre que a pesar de que cualesquiera de las dos variables son independientes, es decir, sus factores función de distribución marginal, las tres no son independientes.

Respuestas

2.38.

x	0	1	2	3
$f(x)$	1/8	3/8	3/8	1/8

2.39.

x	0	1	2
$f(x)$	3/28	15/28	5/14

2.40.

x	0	1	2
$f(x)$	9/64	15/32	25/64

2.42.

x	0	1	2	3	4
$f(x)$	$\frac{194{,}580}{270{,}725}$	$\frac{69{,}184}{270{,}725}$	$\frac{6768}{270{,}725}$	$\frac{192}{270{,}725}$	$\frac{1}{270{,}725}$

2.43.

x	0	1	2	3
$F(x)$	1/8	1/2	7/8	1

2.46. (*a*)

x	1	2	3	4
$f(x)$	1/8	1/4	3/8	1/4

(*b*) 3/4 (*c*) 7/8 (*d*) 3/8 (*e*) 7/8

2.47. (*a*) 3 (*b*) $e^{-3} - e^{-6}$ (*c*) e^{-9} (*d*) $1 - e^{-3}$

2.48. $F(x) = \begin{cases} 1 - e^{-3x} & x \geqq 0 \\ 0 & x \leqq 0 \end{cases}$

2.49. (*a*) 6/29 (*b*) 15/29 (*c*) 19/116

2.50. $F(x) = \begin{cases} 0 & x \leqq 1 \\ (2x^3 - 2)/29 & 1 \leqq x \leqq 2 \\ (3x^2 + 2)/29 & 2 \leqq x \leqq 3 \\ 1 & x \geqq 3 \end{cases}$

2.51. (*a*) 1/27 (*b*) $f(x) = \begin{cases} x^2/9 & 0 \leqq x < 3 \\ 0 & \text{de otra manera} \end{cases}$ (*c*) 26/27 (*d*) 7/27

2.53. (*a*) 1/2 (*b*) 1/2 (*c*) 3/4 (*d*) $F(x) = \begin{cases} 0 & x \leqq 0 \\ x^2/4 & 0 \leqq x \leqq 2 \\ 1 & x \geqq 2 \end{cases}$

2.54. (*a*) 1/36 (*b*) 1/6 (*c*) 1/4 (*d*) 5/6 (*e*) 1/6 (*f*) 1/6 (*g*) 1/2

2.55. (*a*) $f_1(x) = \begin{cases} x/6 & x = 1, 2, 3 \\ 0 & \text{otro } x \end{cases}$ (*b*) $f_2(y) = \begin{cases} y/6 & y = 1, 2, 3 \\ 0 & \text{otro } y \end{cases}$

2.56. (*a*) 3/2 (*b*) 1/4 (*c*) 29/64 (*d*) 5/16

2.57. (*a*) $F_1(x) = \begin{cases} 0 & x \leqq 0 \\ \frac{1}{2}(x^3 + x) & 0 \leqq x \leqq 1 \\ 1 & x \geqq 1 \end{cases}$ (*b*) $F_2(y) = \begin{cases} 0 & y \leqq 0 \\ \frac{1}{2}(y^3 + y) & 0 \leqq y \leqq 1 \\ 1 & y \geqq 1 \end{cases}$

2.58. (*a*) $f(x \mid y) = f_1(x)$ para $y = 1, 2, 3$ (*véase* problema 2.55)

(*b*) $f(y \mid x) = f_2(y)$ para $x = 1, 2, 3$ (*véase* problema 2.55)

2.59. (*a*) $f(x \mid y) = \begin{cases} (x+y)/(y+\frac{1}{2}) & 0 \leqq x \leqq 1,\ 0 \leqq y \leqq 1 \\ 0 & \text{otro } x,\ \ 0 \leqq y \leqq 1 \end{cases}$

(*b*) $f(y \mid x) = \begin{cases} (x+y)/(x+\frac{1}{2}) & 0 \leqq x \leqq 1,\ 0 \leqq y \leqq 1 \\ 0 & 0 \leqq x \leqq 1,\ \text{otro } y \end{cases}$

2.60. (*a*) $f(x \mid y) = \begin{cases} (x^2+y^2)/(y^2+\frac{1}{3}) & 0 \leqq x \leqq 1,\ 0 \leqq y \leqq 1 \\ 0 & \text{otro } x,\ \ 0 \leqq y \leqq 1 \end{cases}$

(*b*) $f(y \mid x) = \begin{cases} (x^2+y^2)/(x^2+\frac{1}{3}) & 0 \leqq x \leqq 1,\ 0 \leqq y \leqq 1 \\ 0 & 0 \leqq x \leqq 1,\ \text{otro } y \end{cases}$

2.61. (*a*) $f(x \mid y) = \begin{cases} e^{-x} & x \geqq 0,\ y \geqq 0 \\ 0 & x < 0,\ y \geqq 0 \end{cases}$ (*b*) $f(y \mid x) = \begin{cases} e^{-y} & x \geqq 0,\ y \geqq 0 \\ 0 & x \geqq 0,\ y < 0 \end{cases}$

2.62. $e^{-\sqrt{y}}/2\sqrt{y}$ para $y > 0$; 0 de otra manera

2.64. $(2\pi)^{-1/2} y^{-1/2} e^{-y/2}$ para $y > 0$; 0 de otra manera

2.66. $1/\pi$ para $-\pi/2 < y < \pi/2$; 0 de otra manera

2.68. (*a*) $g(y) = \begin{cases} \frac{1}{6} & -5 < y < 1 \\ 0 & \text{de otra manera} \end{cases}$ (*b*) $g(y) = \begin{cases} \frac{1}{6}(1-y)^{-2/3} & 0 < y < 1 \\ \frac{1}{6}(y-1)^{-2/3} & 1 < y < 2 \\ 0 & \text{de otra manera} \end{cases}$

2.70. $ve^{-v}/(1+u)^2$ para $u \geqq 0,\ v \geqq 0$; 0 de otra manera

2.73. $g(z) = \begin{cases} -\ln z & 0 < z < 1 \\ 0 & \text{de otra manera} \end{cases}$

2.74. $g(u) = \begin{cases} u & 0 \leqq u \leqq 1 \\ 2-u & 1 \leqq u \leqq 2 \\ 0 & \text{de otra manera} \end{cases}$

2.75. $g(u) = \begin{cases} ue^{-u} & u \geqq 0 \\ 0 & u < 0 \end{cases}$

2.77. $g(x) = \begin{cases} x^3 e^{-x}/6 & x \geqq 0 \\ 0 & x < 0 \end{cases}$

2.78. 1/4

2.79. 61/72

2.81. (*a*) 2 (*b*) $F(x) = \begin{cases} 0 & x < 1 \\ 1 - 3^{-y} & y \leqq x < y+1;\ y = 1, 2, 3, \ldots \end{cases}$ (*d*) 26/81 (*e*) 1/9

2.82. (*a*) 4 (*b*) $F(x) = \begin{cases} 1 - e^{-2x}(2x+1) & x \geqq 0 \\ 0 & x < 0 \end{cases}$ (*d*) $3e^{-2}$ (*e*) $5e^{-4} - 7e^{-6}$

2.83. (*a*) 3/7 (*b*) 5/7

2.84. (*a*) $c = 1$ (*b*) $e^{-4} - 3e^{-2} + 2e^{-1}$

2.86. (*a*) $c_1 = 2,\ c_2 = 9$ (*b*) $9e^{-2} - 14e^{-3}$ (*c*) $4e^{-5} - 4e^{-7}$ (*d*) $e^{-2} - e^{-4}$ (*e*) $4e^{-3}$

2.88. (*a*) 1/4 (*b*) 27/64 (*c*) $f_1(x) = \begin{cases} x + \frac{1}{2} & 0 < x < 1 \\ 0 & \text{de otra manera} \end{cases}$ (*d*) $f_2(y) = \begin{cases} \frac{1}{4}(y+1) & 0 < y < 2 \\ 0 & \text{de otra manera} \end{cases}$

2.90. (*a*) $\begin{cases} e^{-2y}/\sqrt{y} & y > 0 \\ 0 & \text{de otra manera} \end{cases}$ (*b*) $\begin{cases} 18e^{-2u} & u > 0 \\ 0 & \text{de otra manera} \end{cases}$

2.91. (*b*) $\frac{1}{2}(1 - \ln 2)$ (*c*) $\frac{1}{6} + \frac{1}{2}\ln 2$ (*d*) $\frac{1}{2}\ln 2$

2.93. $g(z) = \begin{cases} \frac{1}{2}e^{-z/2} & z \geqq 0 \\ 0 & z < 0 \end{cases}$

2.94. (*b*) 7/18

2.95. (*b*) 15/256 (*c*) 9/16 (*d*) 0

2.100. (*a*) 45/512 (*b*) 1/14

2.102. $\sqrt{2}/2$

CAPÍTULO 3

Esperanza matemática

DEFINICIÓN DE ESPERANZA MATEMÁTICA

Un concepto muy importante en probabilidad y estadística es el de *esperanza matemática*, *valor esperado*, o simplemente, *esperanza* de una variable aleatoria. Para una variable aleatoria discreta X, que tiene los posibles valores $x_1, \ldots, x_n$, la esperanza de X se define como

$$E(X) = x_1 P(X = x_1) + \cdots + x_n P(X = x_n) = \sum_{j=1}^{n} x_j P(X = x_j) \tag{1}$$

o de manera equivalente, si $P(X = x_j) = f(x_j)$,

$$E(X) = x_1 f(x_1) + \cdots + x_n f(x_n) = \sum_{j=1}^{n} x_j f(x_j) = \sum xf(x) \tag{2}$$

donde la última sumatoria toma todos los valores apropiados de x. Como un caso especial de (2), donde todas las probabilidades son iguales, tenemos

$$E(X) = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{3}$$

la cual recibe el nombre de *media aritmética*, o simplemente *media* de $x_1, x_2, \ldots, x_n$.

Si X toma un número infinito de valores $x_1, x_2, \ldots$, entonces $E(X) = \sum_{j=1}^{\infty} x_j f(x_j)$ siempre y cuando las series infinitas converjan absolutamente.

Para una variable aleatoria continua X con función de densidad $f(x)$, la esperanza de X se define como

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx \tag{4}$$

siempre y cuando la integral converja absolutamente.

La esperanza de X se denomina con frecuencia la *media* de X y se denota por μ_X, o simplemente μ, cuando se sobreentiende de cuál variable aleatoria se trata.

La media, o esperanza de X, da un valor que es representativo o promedio de los valores de X, y por esta razón con frecuencia se denomina la *medida de tendencia central*. En la página 91 se tratan otras medidas.

EJEMPLO 3.1. Suponga que se va a jugar una partida con un dado, que se supone está balanceado. En este juego un jugador gana \$20 si obtiene 2; gana \$40 si obtiene 4; pierde \$30 si obtiene 6; y no gana ni pierde si obtiene cualquier otra cara del dado. Encuentre la suma esperada de dinero que él puede ganar.

Sea X la variable aleatoria que indica la cantidad de dinero ganada en cualquier lanzamiento. Las cantidades que es posible ganar cuando la cara del dado cae en 1, 2, . . ., 6 son $x_1, x_2, \ldots, x_6$, respectivamente, mientras que sus probabilidades son $f(x_1), f(x_2), \ldots, f(x_6)$. La función de probabilidad para X se muestra en la tabla 3-1. Por consiguiente, el valor esperado o esperanza es

$$E(X) = (0)\left(\frac{1}{6}\right) + (20)\left(\frac{1}{6}\right) + (0)\left(\frac{1}{6}\right) + (40)\left(\frac{1}{6}\right) + (0)\left(\frac{1}{6}\right) + (-30)\left(\frac{1}{6}\right) = 5$$

Tabla 3-1

x_j	0	+20	0	+40	0	−30
$f(x_j)$	1/6	1/6	1/6	1/6	1/6	1/6

Se deduce que el jugador puede esperar ganar \$5. Por consiguiente, en un juego justo se espera que el jugador pague \$5 por participar.

EJEMPLO 3.2. La función de densidad de una variable aleatoria X está dada por

$$f(x) = \begin{cases} \frac{1}{2}x & 0 < x < 2 \\ 0 & \text{de otra manera} \end{cases}$$

El valor esperado de X es entonces

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx = \int_0^2 x\left(\frac{1}{2}x\right)dx = \int_0^2 \frac{x^2}{2}\,dx = \frac{x^3}{6}\Big|_0^2 = \frac{4}{3}$$

FUNCIONES DE VARIABLES ALEATORIAS

Sea X una variable aleatoria discreta con función de probabilidad $f(x)$. Entonces $Y = g(X)$ también es una variable aleatoria discreta, y la función de probabilidad de Y es

$$h(y) = P(Y = y) = \sum_{\{x \mid g(x)=y\}} P(X = x) = \sum_{\{x \mid g(x)=y\}} f(x)$$

Si X toma los valores $x_1, x_2, \ldots, x_n$ y Y toma los valores $y_1, y_2, \ldots, y_m$ $(m \leq n)$, entonces $y_1h(y_1) + y_2h(y_2) + \cdots + y_mh(y_m) = g(x_1)f(x_1) + g(x_2)f(x_2) + \cdots + g(x_n)f(x_n)$.
Por consiguiente,

$$\begin{aligned} E[g(X)] &= g(x_1)f(x_1) + g(x_2)f(x_2) + \cdots + g(x_n)f(x_n) \\ &= \sum_{j=1}^{n} g(x_j)f(x_j) = \sum g(x)f(x) \end{aligned} \tag{5}$$

De igual manera, si X es una variable aleatoria continua con densidad de probabilidad $f(x)$, entonces se puede demostrar que

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,dx \tag{6}$$

Observe que (5) y (6) no involucran, respectivamente, la función de probabilidad y la función de densidad de probabilidad de $Y = g(X)$.

Es fácil hacer generalizaciones para funciones de dos o más variables aleatorias. Por ejemplo, si X y Y son variables aleatorias continuas con función de densidad conjunta $f(x, y)$, entonces la esperanza de $g(X, Y)$ está dada por

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y)\, dx\, dy \tag{7}$$

EJEMPLO 3.3. Si X es la variable aleatoria del ejemplo 3.2,

$$E(3X^2 - 2X) = \int_{-\infty}^{\infty} (3x^2 - 2x) f(x)\, dx = \int_{0}^{2} (3x^2 - 2x)\left(\frac{1}{2}x\right) dx = \frac{10}{3}$$

ALGUNOS TEOREMAS SOBRE LA ESPERANZA

***Teorema 3-1*:** Si c es una constante, entonces

$$E(cX) = cE(X) \tag{8}$$

***Teorema 3-2*:** Si X y Y son dos variables aleatorias cualesquiera, entonces

$$E(X + Y) = E(X) + E(Y) \tag{9}$$

***Teorema 3-3*:** Si X y Y son variables aleatorias independientes, entonces

$$E(XY) = E(X)E(Y) \tag{10}$$

Fácilmente se pueden hacer generalizaciones de estos teoremas.

LA VARIANZA Y LA DESVIACIÓN ESTÁNDAR

Observamos, en la página 82, que la esperanza de una variable aleatoria X con frecuencia se denomina *media* y se denota como μ. Otro valor de gran importancia en probabilidad y estadística es la *varianza*, que se define así

$$\text{Var}\,(X) = E[(X - \mu)^2] \tag{11}$$

La varianza es un número positivo. La raíz cuadrada positiva de la varianza se llama *desviación estándar*, y está dada por

$$\sigma_X = \sqrt{\text{Var}\,(X)} = \sqrt{E[(X - \mu)^2]} \tag{12}$$

En donde no pueda haber confusión, la desviación estándar con frecuencia se denomina σ, en lugar de σ_x, y en tal caso la varianza es σ^2.

Si X es una variable aleatoria discreta que toma los valores $x_1, x_2, \ldots, x_n$, y tiene función de probabilidad $f(x)$, entonces la varianza está dada por

$$\sigma_X^2 = E[(X - \mu)^2] = \sum_{j=1}^{n} (x_j - \mu)^2 f(x_j) = \sum (x - \mu)^2 f(x) \tag{13}$$

En el caso especial de (*13*) donde las probabilidades son iguales, tenemos

$$\sigma^2 = [(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2]/n \tag{14}$$

la cual es la varianza para un conjunto de n números $x_1, \ldots, x_n$.

Si X toma un número infinito de valores $x_1, x_2, \ldots$, entonces $\sigma_X^2 = \sum_{j=1}^{\infty} (x_j - \mu)^2 f(x_j)$, siempre y cuando la serie converja.

Si X es una variable aleatoria continua con función de densidad $f(x)$, entonces la varianza está dada por

$$\sigma_X^2 = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx \tag{15}$$

siempre y cuando la integral converja.

La varianza (o la desviación estándar) es una medida de la *dispersión*, o de la *variación*, de los valores de la variable aleatoria alrededor de la media μ. Si los valores tienden a concentrarse cerca de la media, la varianza es pequeña, mientras que si los valores tienden a distribuirse lejos de la media, la varianza es grande. La situación se muestra de manera gráfica en la figura 3-1 para el caso de dos distribuciones continuas que tienen la misma media μ.

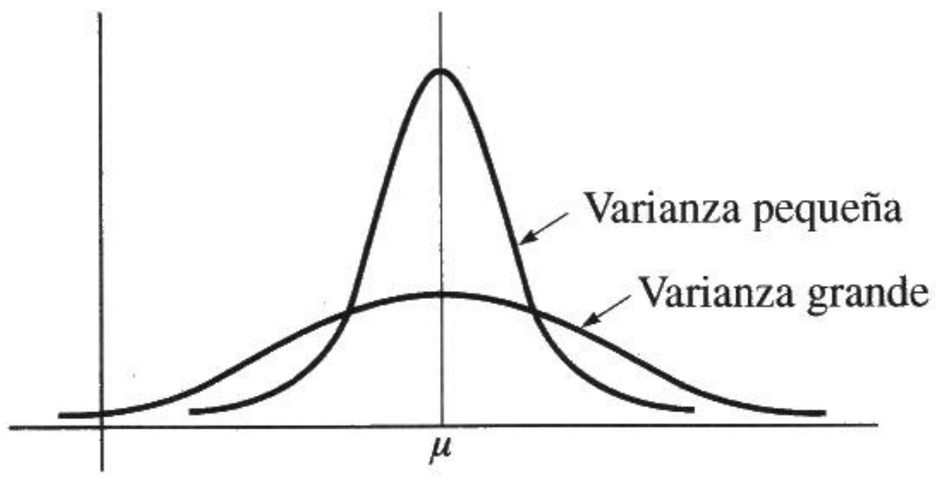


Fig. 3-1

EJEMPLO 3.4. Encuentre la varianza y la desviación estándar de la variable aleatoria del ejemplo 3.2.

Como se encontró en el ejemplo 3.2, la media es $\mu = E(X) = 4/3$. Entonces la varianza está dada por

$$\sigma^2 = E\left[\left(X-\frac{4}{3}\right)^2\right] = \int_{-\infty}^{\infty}\left(x-\frac{4}{3}\right)^2 f(x)\,dx = \int_0^2 \left(x-\frac{4}{3}\right)^2\left(\frac{1}{2}x\right)dx = \frac{2}{9}$$

y así la desviación estándar es $\sigma = \sqrt{\frac{2}{9}} = \frac{\sqrt{2}}{3}$

Observe que si X tiene ciertas *dimensiones* o *unidades*, como *centímetros* (cm), entonces la varianza de X tiene unidades en cm^2, mientras que la desviación estándar tiene las mismas unidades que X, es decir, centímetros. Es por esta razón que con frecuencia se utiliza la desviación estándar.

ALGUNOS TEOREMAS SOBRE LA VARIANZA

***Teorema 3-4*:** $$\sigma^2 = E[(X-\mu)^2] = E(X^2) - \mu^2 = E(X^2) - [E(X)]^2 \tag{16}$$

donde $\mu = E(X)$

***Teorema 3-5*:** Si c es cualquier constante,

$$\text{Var}\,(cX) = c^2\,\text{Var}\,(X) \tag{17}$$

***Teorema 3-6*:** La cantidad $E[(X-a)^2]$ es mínima cuando $a = \mu = E(X)$.

***Teorema 3-7*:** Si X y Y son variables aleatorias independientes,

$$\text{Var}\,(X+Y) = \text{Var}\,(X) + \text{Var}\,(Y) \quad \text{o} \quad \sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 \tag{18}$$

$$\text{Var}\,(X-Y) = \text{Var}\,(X) + \text{Var}\,(Y) \quad \text{o} \quad \sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 \tag{19}$$

Se pueden hacer fácilmente generalizaciones del teorema 3-7 con más de dos variables independientes. En palabras, la varianza de la suma de variables independientes es igual a la suma de sus varianzas.

VARIABLES ALEATORIAS ESTANDARIZADAS

Sea X una variable aleatoria con media μ y desviación estándar $\sigma(\sigma > 0)$. Entonces podemos definir una *variable aleatoria estandarizada* asociada dada por

$$X^* = \frac{X - \mu}{\sigma} \tag{20}$$

Una propiedad importante de X^* es que tiene media de cero y varianza de 1, lo cual justifica el nombre de *estandarizada*, es decir,

$$E(X^*) = 0, \qquad \text{Var}\,(X^*) = 1 \tag{21}$$

En algunas ocasiones, los valores de una variable estandarizada se denominan *registros estándar*, y se dice entonces que X se expresa en *unidades estándar* (es decir, σ se toma como la unidad al medir $X - \mu$).

Las variables estandarizadas son útiles para comparar distribuciones diferentes.

MOMENTOS

El *momento r-ésimo de una variable aleatoria X alrededor de la media* μ, también llamado el *momento central r-ésimo*, se define como

$$\mu_r = E[(X - \mu)^r] \tag{22}$$

donde $r = 0, 1, 2, \ldots$. Se deduce que $\mu_0 = 1$, $\mu_1 = 0$, y $\mu_2 = \sigma^2$, es decir, el segundo momento central o segundo momento alrededor de la media es la varianza. Si suponemos la convergencia absoluta, tenemos

$$\mu_r = \sum (x - \mu)^r f(x) \quad \text{(variable discreta)} \tag{23}$$

$$\mu_r = \int_{-\infty}^{\infty} (x - \mu)^r f(x)\, dx \quad \text{(variable continua)} \tag{24}$$

El *momento r-ésimo de X alrededor del origen*, también llamado el *momento central r-ésimo*, se define como

$$\mu_r' = E(X^r) \tag{25}$$

donde $r = 0, 1, 2, \ldots$, y en este caso hay fórmulas análogas a (*23*) y (*24*) en las cuales $\mu = 0$.

La relación que existe entre estos momentos está dada por

$$\mu_r = \mu_r' - \binom{r}{1}\mu_{r-1}'\mu + \cdots + (-1)^j\binom{r}{j}\mu_{r-j}'\mu^j + \cdots + (-1)^r\mu_0'\mu^r \tag{26}$$

Como casos especiales tenemos, usando $\mu_1' = \mu$ y $\mu_0' = 1$,

$$\mu_2 = \mu_2' - \mu^2$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu + 2\mu^3 \tag{27}$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu + 6\mu_2'\mu^2 - 3\mu^4$$

FUNCIONES GENERADORAS DE MOMENTO

La *función generadora de momento* de X se define como

$$M_X(t) = E(e^{tX}) \tag{28}$$

esto es, suponiendo que hay convergencia,

$$M_X(t) = \sum e^{tx} f(x) \text{ (variable discreta)} \tag{29}$$

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx \quad \text{(variable continua)} \tag{30}$$

Podemos demostrar que la expansión de las series de Taylor es [problema 3.15(a)]

$$M_X(t) = 1 + \mu t + \mu_2' \frac{t^2}{2!} + \cdots + \mu_r' \frac{t^r}{r!} + \cdots \tag{31}$$

Dado que los coeficientes de la expansión nos permiten encontrar los momentos, la razón del nombre *función generadora de momento* es aparente. A partir de la expansión podemos demostrar que [Problema 3.15(b)]

$$\mu_r' = \frac{d^r}{dt^r} M_X(t)\bigg|_{t=0} \tag{32}$$

es decir, μ_r' es la r-ésima derivada de $M_X(t)$ evaluada en $t = 0$. En donde no puede haber confusión, escribimos con frecuencia $M(t)$ en lugar de $M_X(t)$.

ALGUNOS TEOREMAS SOBRE FUNCIONES GENERADORAS DE MOMENTO

Teorema 3-8: Si $M_x(t)$ es la función generadora de momento de una variable aleatoria X y a y b ($b \neq 0$) son constantes, entonces la función generadora de momento de $(X + a)/b$ es

$$M_{(X+a)/b}(t) = e^{at/b} M_X\left(\frac{t}{b}\right) \tag{33}$$

Teorema 3-9: Si X y Y son variables aleatorias independientes con funciones generadoras de momento de $M_X(t)$ y de $M_Y(t)$, respectivamente, entonces

$$M_{X+Y}(t) = M_X(t) M_Y(t) \tag{34}$$

Se pueden hacer fácilmente generalizaciones del teorema 3-9 para más de dos variables aleatorias independientes. En palabras, la función generadora de momento de la suma de variables aleatorias independientes es igual al producto de sus funciones generadoras de momento.

Teorema 3-10 **(Teorema de la unicidad):** Supongamos que X y Y son variables aleatorias que tienen funciones generadoras de momento $M_X(t)$ y $M_Y(t)$, respectivamente. Entonces X y Y tienen la misma distribución de probabilidad si y sólo si $M_X(t) = M_Y(t)$, de manera idéntica.

FUNCIONES CARACTERÍSTICAS

Si remplazamos $t = i\omega$, donde i es la unidad imaginaria en la función generadora de momento, obtenemos una función importante, llamada la *función característica*. Lo denotamos como

$$\phi_X(\omega) = M_X(i\omega) = E(e^{i\omega X}) \tag{35}$$

Se deduce que

$$\phi_X(\omega) = \sum e^{i\omega x} f(x) \quad \text{(variable discreta)} \tag{36}$$

$$\phi_X(\omega) = \int_{-\infty}^{\infty} e^{i\omega x} f(x)\, dx \quad \text{(variable continua)} \tag{37}$$

Dado que $|e^{i\omega x}| = 1$, las series y la integral siempre covergen de manera absoluta.

Los resultados correspondientes (*31*) y (*32*) son

$$\phi_X(\omega) = 1 + i\mu\omega - \mu_2' \frac{\omega^2}{2!} + \cdots + i^r \mu_r' \frac{\omega^r}{r!} + \cdots \tag{38}$$

donde

$$\mu_r' = (-1)^r i^r \frac{d^r}{d\omega^r} \phi_X(\omega)\bigg|_{\omega=0} \tag{39}$$

Donde no pueda haber confusiones, con frecuencia escribimos $\phi(\omega)$ en lugar de $\phi_X(\omega)$.

Los teoremas para funciones características correspondientes a los teoremas 3-8, 3-9, y 3-10 son los siguientes.

Teorema 3-11: Si $\phi(\omega)$ es la función característica de la variable aleatoria X y a y $b(b \neq 0)$ son constantes, entonces la función característica de $(X + a)/b$ es

$$\phi_{(X+a)/b}(\omega) = e^{ai\omega/b}\phi_X\left(\frac{\omega}{b}\right) \tag{40}$$

Teorema 3-12: Si X y Y son variables aleatorias independientes con funciones características $\phi_X(\omega)$ y $\phi_Y(\omega)$, respectivamente, entonces

$$\phi_{X+Y}(\omega) = \phi_X(\omega)\phi_Y(\omega) \tag{41}$$

De manera más general, la función característica de una suma de variables aleatorias independientes es igual al producto de sus funciones características.

Teorema 3-13 **(Teorema de la unicidad):** Suponga que X y Y son variables aleatorias que tienen funciones características $\phi_X(\omega)$ y $\phi_Y(\omega)$, respectivamente. Entonces X y Y tienen la misma distribución de probabilidad si y sólo si $\phi_X(\omega) = \phi_Y(\omega)$, de manera idéntica.

Una razón importante para introducir la función característica es que (*37*) representa la *transformada de Fourier* de la función de densidad $f(x)$. A partir de la teoría de las transformadas de Fourier, podemos determinar fácilmente la función de densidad a partir de la función característica. De hecho,

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\omega x}\phi_X(\omega)\,d\omega \tag{42}$$

la cual, con frecuencia, se denomina *fórmula de inversión* o *transformada inversa de Fourier*. De manera similar podemos demostrar en el caso discreto, que la función de probabilidad $f(x)$ se puede obtener a partir de (*36*) usando la *serie de Fourier*, la cual es análoga a la integral de Fourier para el caso discreto. *Véase* el problema 3.39.

Otra razón para usar la función característica es que ésta siempre existe, mientras que la función generadora de momento puede no existir.

VARIANZA PARA DISTRIBUCIONES CONJUNTAS. COVARIANZA

Los resultados dados anteriormente pueden extenderse a dos o más variables. Por ejemplo, si X y Y son dos variables aleatorias continuas con función de densidad conjunta $f(x, y)$, las medias, o esperanzas, de X y Y son

$$\mu_X = E(X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf(x,y)\,dx\,dy, \qquad \mu_Y = E(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf(x,y)\,dx\,dy \tag{43}$$

y las varianzas son

$$\begin{aligned}\sigma_X^2 &= E[(X-\mu_X)^2] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)^2 f(x,y)\,dx\,dy \\ \sigma_Y^2 &= E[(Y-\mu_Y)^2] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y-\mu_Y)^2 f(x,y)\,dx\,dy\end{aligned} \tag{44}$$

Observe que las funciones de densidad marginal de X y Y no participan directamente en (*43*) y (*44*).

La *covarianza* es otra cantidad que surge en el caso de dos variables X y Y, y se define por

$$\sigma_{XY} = \mathrm{Cov}\,(X, Y) = E[(X-\mu_X)(Y-\mu_Y)] \tag{45}$$

En términos de la función de densidad conjunta $f(x, y)$, tenemos

$$\sigma_{XY} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)f(x,y)\,dx\,dy \tag{46}$$

De manera similar se pueden hacer observaciones para dos variables aleatorias discretas. En tales casos, (*43*) y (*46*) se remplazan por

$$\mu_X = \sum_x \sum_y xf(x,y) \quad ; \quad \mu_Y = \sum_x \sum_y yf(x,y) \tag{47}$$

$$\sigma_{XY} = \sum_x \sum_y (x-\mu_X)(y-\mu_Y)f(x,y) \tag{48}$$

donde las sumatorias toman todos los valores discretos de X y Y.

Los siguientes son algunos teoremas importantes en covarianza.

***Teorema 3-14*:** $$\sigma_{XY} = E(XY) - E(X)E(Y) = E(XY) - \mu_X\mu_Y \tag{49}$$

***Teorema 3-15*:** Si X y Y son variables aleatorias independientes, entonces

$$\sigma_{XY} = \text{Cov}(X, Y) = 0 \tag{50}$$

***Teorema 3-16*:** $$\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y) \pm 2\,\text{Cov}(X, Y) \tag{51}$$

o $$\sigma^2_{X\pm Y} = \sigma^2_X + \sigma^2_Y \pm 2\sigma_{XY} \tag{52}$$

***Teorema 3-17*:** $$|\sigma_{XY}| \leqq \sigma_X\sigma_Y \tag{53}$$

El inverso del teorema 3-15 no necesariamente es cierto. Si X y Y son independientes, el teorema 3-16 se reduce al teorema 3-7.

COEFICIENTE DE CORRELACIÓN

Si X y Y son independientes, entonces $\text{Cov}(X, Y) = \sigma_{XY}$. Por otro lado, si X y Y son completamente dependientes, por ejemplo, cuando $X = Y$, entonces $\text{Cov}(X, Y) = \sigma_{XY} = \sigma_X\,\sigma_Y$. A partir de este razonamiento llegamos a la medida de la dependencia de las variables X y Y, dada por

$$\rho = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} \tag{54}$$

Podemos llamar ρ como el *coeficiente de correlación*. Del teorema 3-17 vemos que $-1 \leqq \rho \leqq 1$. En el caso donde $\rho = 0$ (es decir, la covarianza es cero), decimos que las variables X y Y *no están correlacionadas*. Sin embargo, en tales casos las variables pueden o no ser independientes. En el capítulo 8 ampliaremos el análisis sobre los casos de correlación.

ESPERANZA CONDICIONAL, VARIANZA Y MOMENTOS

Si X y Y tienen función de densidad conjunta $f(x, y)$, entonces, según el capítulo 2, la función de densidad condicional de Y, dado X, *es* $f(x \mid y) = f(x, y)/f_1(x)$ donde $f_1(x)$ es la función de densidad marginal de X. Podemos definir la *esperanza condicional*, o *media condicional*, de Y, dado X, por

$$E(Y \mid X = x) = \int_{-\infty}^{\infty} yf(y \mid x)\,dy \tag{55}$$

donde en el caso continuo "$X = x$" se interpreta como $x < X \leqq x + dx$. Los teoremas 3-1 y 3-2 también son válidos para la esperanza condicional.

Podemos observar las siguientes propiedades:

1. $E(Y \mid X = x) = E(Y)$ cuando X y Y son independientes.
2. $E(Y) = \int_{-\infty}^{\infty} E(Y \mid X = x)f_1(x)\,dx.$

Conviene, con frecuencia, calcular las esperanzas usando la propiedad 2, en lugar de hacerlo directamente.

EJEMPLO 3.5. El tiempo promedio de viaje a una ciudad distante es c horas en autobús. Una mujer no puede decidir si viaja en su automóvil o si toma el autobús, así que decide lanzar una moneda. ¿Cuál es su tiempo esperado de viaje?

En este caso estamos tratando con la distribución conjunta del resultado del lanzamiento de una moneda, X, y el tiempo de viaje, Y, donde $Y = Y_{\text{car}}$ si $X = 0$ y $Y = Y_{\text{bus}}$ si $X = 1$. Presumiblemente tanto Y_{car} como Y_{bus} son independientes de X, de manera que según la propiedad 1 anterior

$$E(Y \mid X = 0) = E(Y_{\text{car}} \mid X = 0) = E(Y_{\text{car}}) = c$$

y

$$E(Y \mid X = 1) = E(Y_{\text{bus}} \mid X = 1) = E(Y_{\text{bus}}) = b$$

Entonces, para una moneda balanceada, obtenemos en la propiedad 2 (en la que remplazamos la integral por una sumatoria),

$$E(Y) = E(Y \mid X = 0)P(X = 0) + E(Y \mid X = 1)P(X = 1) = \frac{c + b}{2}$$

De manera similar podemos definir la varianza condicional de Y, dado X, como

$$E[(Y - \mu_2)^2 \mid X = x] = \int_{-\infty}^{\infty} (y - \mu_2)^2 f(y \mid x)\, dy \tag{56}$$

donde $\mu_2 = E(Y \mid X = x)$. También podemos definir el momento r-ésimo de Y alrededor de cualquier valor a, dado X, como

$$E[(Y - a)^r \mid X = x] = \int_{-\infty}^{\infty} (y - a)^r f(y \mid x)\, dy \tag{57}$$

Los teoremas para varianza y momentos se extienden a varianza y momentos condicionales.

DESIGUALDAD DE CHEBYSHEV

Un teorema importante en probabilidad y estadística que revela la propiedad general de las variables aleatorias continuas o discretas con media finita y varianza, se conoce como la *desigualdad de Chebyshev*.

Teorema 3-18 **(desigualdad de Chebyshev):** Supongamos que X es una variable aleatoria (discreta o continua) con media μ y varianza σ^2, las cuales son finitas. Entonces si ϵ es cualquier número positivo,

$$P(|X - \mu| \geqq \epsilon) \leqq \frac{\sigma^2}{\epsilon^2} \tag{58}$$

o, con $\epsilon = k\sigma$,

$$P(|X - \mu| \geqq k\sigma) \leqq \frac{1}{k^2} \tag{59}$$

EJEMPLO 3.6. Remplazando $k = 2$ en la desigualdad de Chebyshev (*59*), vemos que

$$P(|X - \mu| \geqq 2\sigma) \leqq 0.25 \qquad \text{o} \qquad P(|X - \mu| < 2\sigma) \geqq 0.75$$

En palabras, la probabilidad de que X difiera de su media por más de 2 desviaciones estándar es menor o igual que 0.25; de igual modo, la probabilidad de que X esté entre 2 desviaciones estándar de su media es mayor o igual que 0.75. Esto es bastante asombroso en vista de que ni siquiera hemos especificado la distribución de probabilidad de X.

LEY DE LOS GRANDES NÚMEROS

El siguiente teorema, llamado *ley de los grandes números*, es una consecuencia interesante de la desigualdad de Chebyshev.

Teorema 3-19 **(Ley de los grandes números):** Sean $X_1, X_2, \ldots, X_n$ variables aleatorias mutuamente independientes (discretas o continuas), cada una con media μ finita y varianza σ^2. Entonces, si $S_n = X_1 + X_2 + \cdots + X_n$ $(n = 1, 2, \ldots)$,

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| \geqq \epsilon\right) = 0 \tag{60}$$

Puesto que S_n/n es la media aritmética de $X_1, \ldots, X_n$, este teorema establece que la probabilidad de la media aritmética S_n/n difiera de su valor esperado μ por más de ϵ, tiende a cero a medida que $n \to \infty$. Un resultado más fuerte, que esperamos sea cierto, es que $\lim_{n\to\infty} S_n/n = \mu$, pero en realidad es falso. Sin embargo, podemos demostrar que $\lim_{n\to\infty} S_n/n = \mu$ *con probabilidad uno*. Este resultado se conoce como la *ley fuerte de los grandes números*, y por contraste el del teorema 3-19 se conoce como la *ley débil de los grandes números*. Cuando se hace referencia sin calificación a la "ley de los grandes números", la ley débil está implicada.

OTRAS MEDIDAS DE LA TENDENCIA CENTRAL

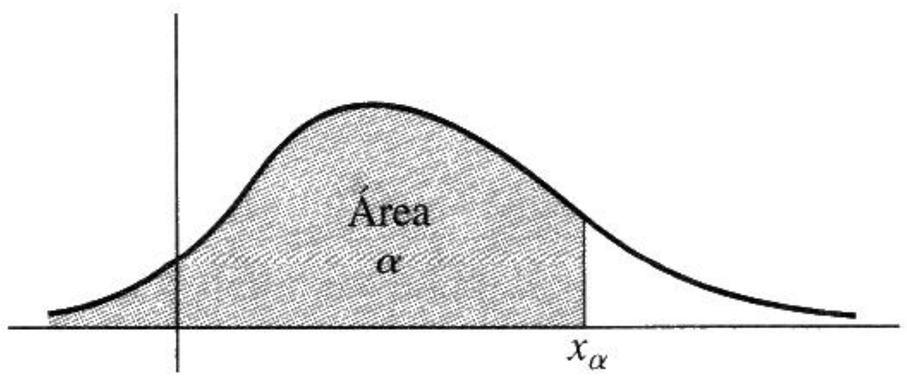
Como ya hemos visto, la media, o esperanza, de una variable aleatoria X provee una medida de la tendencia central para los valores de una distribución. Aunque se utiliza más la media, también se emplean otras dos medidas de la tendencia central. Éstas son la *moda* y la *mediana*.

1. **MODA.** La *moda* de una variable aleatoria discreta es el valor que ocurre con más frecuencia o, en otras palabras, el que tiene la mayor probabilidad de ocurrencia. Algunas veces tenemos dos, tres o más valores que tienen probabilidades relativamente grandes de ocurrencia. En tales casos, decimos que la distribución es *bimodal*, *trimodal* o *multimodal*, respectivamente. La moda de una variable aleatoria continua X es el valor (o valores) de X donde la función de densidad de probabilidad tiene un máximo relativo.

2. **MEDIANA.** La *mediana* es el valor de x para el cual $P(X < x) = \frac{1}{2}$ y $P(X > x) \leq \frac{1}{2}$. En el caso de una distribución continua tenemos $P(X < x) = \frac{1}{2} = P(X > x)$, y la mediana separa la curva de densidad en dos partes con áreas iguales de $1/2$ cada una. En el caso de una distribución discreta, no existe una mediana única (*véase* problema 3.34).

PERCENTILES

Con frecuencia conviene subdividir el área bajo la curva de densidad usando ordenadas, de manera que el área a la izquierda de la ordenada sea algún porcentaje del área total. Los valores correspondientes a tales áreas se llaman *valores percentiles*, o simplemente *percentiles*. Así, por ejemplo, el área a la izquierda de la ordenada x_α en la figura 3-2 es α. Por ejemplo, el área a la izquierda de $x_{0.10}$ sería 0.10 ó 10%, y $x_{0.10}$ podría llamarse *décimo percentil* (también llamado el *primer decil*). La mediana sería el *quincuagésimo percentil* (o *quinto decil*).



OTRAS MEDIDAS DE DISPERSIÓN

Así como hay varias medidas para la tendencia central, además de la media, existen varias medidas de la dispersión o variación de una variable aleatoria, además de la varianza y de la desviación estándar. Algunas de las más comunes son las siguientes.

1. RANGO SEMIINTERCUARTÍLICO. Si $x_{0.25}$ y $x_{0.75}$ representan el 25-avo y 75-avo valores percentiles, la diferencia $x_{0.75} - x_{0.25}$ se conoce como *rango intercuartílico* y $\frac{1}{2}(x_{0.75} - x_{0.25})$ es el *rango semiintercuartílico*.

2. DESVIACIÓN MEDIA. La *desviación media* (DM) de una variable aleatoria X se define como la esperanza de $|X - \mu|$, es decir, asumiendo que hay convergencia,

$$\text{M.D.}(X) = E[|X - \mu|] = \sum |x - \mu| f(x) \qquad \text{(variable discreta)} \tag{61}$$

$$\text{M.D.}(X) = E[|X - \mu|] = \int_{-\infty}^{\infty} |x - \mu| f(x)\, dx \qquad \text{(variable continua)} \tag{62}$$

SESGO Y CURTOSIS

1. SESGO. Con frecuencia una distribución no es simétrica alrededor de ningún valor, pero en lugar de ello tiene una de sus colas más larga que la otra. Si la cola más larga está ubicada a la derecha, como en la figura 3-3, se dice que la distribución es *sesgada hacia la derecha*, mientras que si la cola más larga está a la izquierda, como en la figura 3-4, se dice que está *sesgada hacia la izquierda*. Las medidas que describen esta asimetría se denominan *coeficientes de sesgo*, o simplemente *sesgo*. Una de dichas medidas es

$$\alpha_3 = \frac{E[(X - \mu)^3]}{\sigma^3} = \frac{\mu_3}{\sigma^3} \tag{63}$$

La medida α_3 será positiva o negativa si la distribución es sesgada a la derecha o a la izquierda, respectivamente. Para una distribución simétrica, $\alpha_3 = 0$.

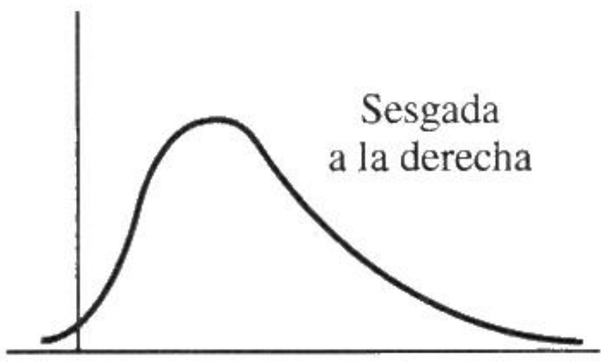


Fig. 3-3

Fig. 3-4

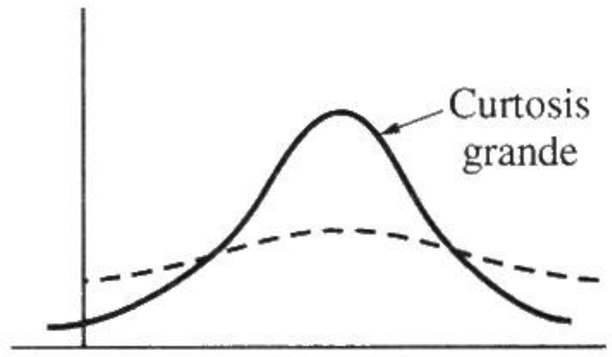


Fig. 3-5

2. CURTOSIS. En algunos casos una distribución puede tener sus valores concentrados cerca de la media, de manera que la distribución tiene un pico grande como se indica en la curva continua de la figura 3-5. En otros casos, la distribución puede ser relativamente plana como se ve en la curva punteada de la figura 3-5. Las medidas que determinan qué tan empinada se encuentra una distribución se denominan *coeficientes de curtosis*, o simplemente *curtosis*. Una medida que se usa con frecuencia está dada por

$$\alpha_4 = \frac{E[(X - \mu)^4]}{\sigma^4} = \frac{\mu_4}{\sigma^4} \tag{64}$$

Generalmente ésta se compara con la curva normal (*véase* capítulo 4), la cual tiene un coeficiente de curtosis igual a 3. *Véase también* el problema 3.41.

Problemas resueltos

ESPERANZA DE VARIABLES ALEATORIAS

3.1. En una lotería hay 200 premios de \$5, 20 premios de \$25 y 5 premios de \$100. Suponiendo que se fabrican y se venden 10,000 boletas, ¿cuál es el precio justo que se debe pagar por cada boleta?

Sea X una variable aleatoria que indica la cantidad de dinero que se gana con una boleta. En la tabla 3-2 se muestran los diferentes valores de X junto con sus probabilidades. Por ejemplo, la probabilidad de obtener una de las 20 boletas que otorga un premio de \$25 es $20/10{,}000 = 0.002$. Así, la esperanza de X en pesos es

$$E(X) = (5)(0.02) + (25)(0.002) + (100)(0.0005) + (0)(0.9775) = 0.2$$

o 20 centavos. Así que el precio justo por cada boleta es de 20 centavos. Sin embargo, puesto que generalmente el propósito de las loterías es ganar dinero, el precio por boleta será más alto.

Tabla 3-2

x (pesos)	5	25	100	0
$P(X = x)$	0.02	0.002	0.0005	0.9775

3.2. Encuentre la esperanza de la suma de puntos al lanzar un par de dados balanceados.

Sean X y Y los puntos que muestran los dos dados. Tenemos

$$E(X) = E(Y) = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + \cdots + 6\left(\frac{1}{6}\right) = \frac{7}{2}$$

Entonces, según el teorema 3-2,

$$E(X + Y) = E(X) + E(Y) = 7$$

3.3. Encuentre la esperanza de una variable aleatoria discreta X cuya función de probabilidad está dada por

$$f(x) = \left(\frac{1}{2}\right)^x \qquad (x = 1, 2, 3, \ldots)$$

Tenemos

$$E(X) = \sum_{x=1}^{\infty} x\left(\frac{1}{2}\right)^x = \frac{1}{2} + 2\left(\frac{1}{4}\right) + 3\left(\frac{1}{8}\right) + \cdots$$

Para resolver esta suma, sea

$$S = \frac{1}{2} + 2\left(\frac{1}{4}\right) + 3\left(\frac{1}{8}\right) + 4\left(\frac{1}{16}\right) + \cdots$$

Entonces

$$\frac{1}{2}S = \quad \frac{1}{4} + 2\left(\frac{1}{8}\right) + 3\left(\frac{1}{16}\right) + \cdots$$

Restando

$$\frac{1}{2}S = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \cdots = 1$$

Por consiguiente, $S = 2$.

3.4. Una variable aleatoria continua X tiene densidad de probabilidad dada por

$$f(x) = \begin{cases} 2e^{-2x} & x > 0 \\ 0 & x \leqq 0 \end{cases}$$

Encuentre (*a*) $E(X)$, (*b*) $E(X^2)$

(*a*)
$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx = \int_{0}^{\infty} x(2e^{-2x})\,dx = 2\int_{0}^{\infty} xe^{-2x}\,dx$$
$$= 2\left[(x)\left(\frac{e^{-2x}}{-2}\right) - (1)\left(\frac{e^{-2x}}{4}\right)\right]\Bigg|_0^{\infty} = \frac{1}{2}$$

(*b*)
$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\,dx = 2\int_{0}^{\infty} x^2 e^{-2x}\,dx$$
$$= 2\left[(x^2)\left(\frac{e^{-2x}}{-2}\right) - (2x)\left(\frac{e^{-2x}}{4}\right) + (2)\left(\frac{e^{-2x}}{-8}\right)\right]\Bigg|_0^{\infty} = \frac{1}{2}$$

3.5. La función de densidad conjunta de dos variables aleatorias X y Y está dada por

$$f(x,y) = \begin{cases} xy/96 & 0 < x < 4,\ 1 < y < 5 \\ 0 & \text{de otra manera} \end{cases}$$

Encuentre (*a*) $E(X)$, (*b*) $E(Y)$, (*c*) $E(XY)$, (*d*) $E(2X + 3Y)$.

(*a*)
$$E(X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf(x,y)\,dx\,dy = \int_{x=0}^{4}\int_{y=1}^{5} x\left(\frac{xy}{96}\right)dx\,dy = \frac{8}{3}$$

(*b*)
$$E(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf(x,y)\,dx\,dy = \int_{x=0}^{4}\int_{y=1}^{5} y\left(\frac{xy}{96}\right)dx\,dy = \frac{31}{9}$$

(*c*)
$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (xy)f(x,y)\,dx\,dy = \int_{x=0}^{4}\int_{y=1}^{5} (xy)\left(\frac{xy}{96}\right)dx\,dy = \frac{248}{27}$$

(*d*)
$$E(2X+3Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (2x+3y)f(x,y)\,dx\,dy = \int_{x=0}^{4}\int_{y=1}^{5} (2x+3y)\left(\frac{xy}{96}\right)dx\,dy = \frac{47}{3}$$

Otro método.

(*c*) Dado que X y Y son independientes, tenemos, usando las partes (*a*) y (*b*),

$$E(XY) = E(X)E(Y) = \left(\frac{8}{3}\right)\left(\frac{31}{9}\right) = \frac{248}{27}$$

(*d*) Según los teoremas 3-1 y 3-2, página 84, junto con (*a*) y (*b*),

$$E(2X+3Y) = 2E(X) + 3E(Y) = 2\left(\frac{8}{3}\right) + 3\left(\frac{31}{9}\right) = \frac{47}{3}$$

3.6. Demuestre el teorema 3-2, página 84.

Sea $f(x, y)$ la función de probabilidad conjunta de X y Y, que se suponen discretas. Entonces,

$$E(X+Y) = \sum_x \sum_y (x+y)f(x,y)$$
$$= \sum_x \sum_y xf(x,y) + \sum_x \sum_y yf(x,y)$$
$$= E(X) + E(Y)$$

Si cualquiera de las variables es continua, la demostración se desarrolla como se hizo anteriormente, remplazando las sumatorias adecuadas por integrales. Observe que el teorema es verdadero tanto si X y Y son independientes, como si no lo son.

3.7. Demuestre el teorema 3-3, página 84.

Sea $f(x, y)$ la función de probabilidad conjunta de X y Y, que se suponen discretas. Si las variables X y Y son independientes, tenemos $f(x, y) = f_1(x)f_2(y)$. Por consiguiente,

$$\begin{aligned} E(XY) &= \sum_x \sum_y xyf(x,y) = \sum_x \sum_y xyf_1(x)f_2(y) \\ &= \sum_x \left[xf_1(x) \sum_y yf_2(y) \right] \\ &= \sum_x [xf_1(x)E(y)] \\ &= E(X)E(Y) \end{aligned}$$

Si cualquier variable es continua, la demostración se desarrolla como se hizo anteriormente, remplazando las sumatorias adecuadas por integrales. Observe que la validez de este teorema gira alrededor de si $f(x, y)$ puede expresarse como una función de x multiplicada por una función de y, para todo x y y, es decir, si X y Y son independientes. Para variables dependientes esto, en general, no es cierto.

VARIANZA Y DESVIACIÓN ESTÁNDAR

3.8. Encuentre (*a*) la varianza, (*b*) la desviación estándar de la suma obtenida al lanzar un par de dados balanceados.

(*a*) En relación con el problema 3.2, tenemos $E(X) = E(Y) = 7/2$. Además,

$$E(X^2) = E(Y^2) = 1^2\left(\frac{1}{6}\right) + 2^2\left(\frac{1}{6}\right) + \cdots + 6^2\left(\frac{1}{6}\right) = \frac{91}{6}$$

Entonces, según el teorema 3-4

$$\mathbf{Var}\,(X) = \mathbf{Var}\,(Y) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$$

y, dado que X y Y son independientes, según el teorema 3-7

$$\mathbf{Var}\,(X + Y) = \mathbf{Var}\,(X) + \mathbf{Var}\,(Y) = \frac{35}{6}$$

(*b*)
$$\sigma_{X+Y} = \sqrt{\mathbf{Var}\,(X + Y)} = \sqrt{\frac{35}{6}}$$

3.9. Encuentre (*a*) la varianza, (*b*) la desviación estándar para la variable aleatoria del problema 3.4.

(*a*) Al igual que en el problema 3.4, la media de X es $\mu = E(X) = \frac{1}{2}$. Entonces la varianza es

$$\begin{aligned} \mathbf{Var}\,(X) = E[(X-\mu)^2] = E\left[\left(X - \frac{1}{2}\right)^2\right] &= \int_{-\infty}^{\infty} \left(x - \frac{1}{2}\right)^2 f(x)\,dx \\ &= \int_0^{\infty} \left(x - \frac{1}{2}\right)^2 (2e^{-2x})\,dx = \frac{1}{4} \end{aligned}$$

Otro método.

Según el teorema 3-4,

$$\mathbf{Var}\,(X) = E[(X-\mu)^2] = E(X^2) - [E(X)]^2 = \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{4}$$

(*b*)
$$\sigma = \sqrt{\text{Var}\,(X)} = \sqrt{\frac{1}{4}} = \frac{1}{2}$$

3.10. Pruebe el teorema 3-4, página 85.

Tenemos

$$\begin{aligned} E[(X-\mu)^2] &= E(X^2 - 2\mu X + \mu^2) = E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2 \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

3.11. Pruebe el teorema 3-6, página 85.

$$\begin{aligned} E[(X-a)^2] &= E[\{(X-\mu) + (\mu - a)\}^2] \\ &= E[(X-\mu)^2 + 2(X-\mu)(\mu-a) + (\mu-a)^2] \\ &= E[(X-\mu)^2] + 2(\mu-a)E(X-\mu) + (\mu-a)^2 \\ &= E[(X-\mu)^2] + (\mu-a)^2 \end{aligned}$$

dado que $E(X-\mu) = E(X) - \mu = 0$. A partir de esto podemos ver que el valor mínimo de $E[(X-a)^2]$ ocurre cuando $(\mu - a)^2 = 0$, es decir, cuando $a = \mu$.

3.12. Si $X^* = (X-\mu)/\sigma$ es una variable aleatoria estandarizada, pruebe que (*a*) $E(X^*) = 0$, (*b*) $\text{Var}(X^*) = 1$.

(*a*)
$$E(X^*) = E\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}\,[E(X-\mu)] = \frac{1}{\sigma}\,[E(X) - \mu] = 0$$

dado que $E(X) = \mu$

(*b*)
$$\text{Var}\,(X^*) = \text{Var}\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2}\,E[(X-\mu)^2] = 1$$

usando el teorema 3-5, página 85, y el hecho de que $E[(X-\mu)^2] = \sigma^2$.

3.13. Pruebe el teorema 3-7, página 85.

$$\begin{aligned} \text{Var}\,(X+Y) &= E[\{(X+Y) - (\mu_X + \mu_Y)\}^2] \\ &= E[\{(X-\mu_X) + (Y-\mu_Y)\}^2] \\ &= E[(X-\mu_X)^2 + 2(X-\mu_X)(Y-\mu_Y) + (Y-\mu_Y)^2] \\ &= E[(X-\mu_X)^2] + 2E[(X-\mu_X)(Y-\mu_Y)] + E[(Y-\mu_Y)^2] \\ &= \text{Var}\,(X) + \text{Var}\,(Y) \end{aligned}$$

usando el hecho de que

$$E[(X-\mu_X)(Y-\mu_Y)] = E(X-\mu_X)E(Y-\mu_Y) = 0$$

dado X y Y, y por consiguiente $X - \mu_X$ y $Y - \mu_Y$, son independientes. La demostración de (*19*), página 85, se da remplazando Y por $-Y$ y usando el teorema 3-5.

MOMENTOS Y FUNCIONES GENERADORAS DE MOMENTOS

3.14. Pruebe el resultado (*26*), página 86.

$$\begin{aligned}\mu_r &= E[(X-\mu)^r] \\ &= E\left[X^r - \binom{r}{1}X^{r-1}\mu + \cdots + (-1)^j\binom{r}{j}X^{r-j}\mu^j \right. \\ &\qquad \left. + \cdots + (-1)^{r-1}\binom{r}{r-1}X\mu^{r-1} + (-1)^r\mu^r\right] \\ &= E(X^r) - \binom{r}{1}E(X^{r-1})\mu + \cdots + (-1)^j\binom{r}{j}E(X^{r-j})\mu^j \\ &\qquad + \cdots + (-1)^{r-1}\binom{r}{r-1}E(X)\mu^{r-1} + (-1)^r\mu^r \\ &= \mu_r' - \binom{r}{1}\mu_{r-1}'\mu + \cdots + (-1)^j\binom{r}{j}\mu_{r-j}'\mu^j \\ &\qquad + \cdots + (-1)^{r-1}r\mu^r + (-1)^{-r}\mu^r\end{aligned}$$

donde los dos últimos términos pueden combinarse para dar $(-1)^{r-1}(r-1)\mu^r$.

3.15. Pruebe (*a*) el resultado (*31*), (*b*) el resultado (*32*), página 87.

(*a*) Usando la expansión en serie de potencia para e^u (3., apéndice A), tenemos

$$\begin{aligned}M_X(t) = E(e^{tX}) &= E\left(1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \cdots\right) \\ &= 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \frac{t^3}{3!}E(X^3) + \cdots \\ &= 1 + \mu t + \mu_2'\frac{t^2}{2!} + \mu_3'\frac{t^3}{3!} + \cdots\end{aligned}$$

(*b*) Esto se deduce inmediatamente del hecho conocido del cálculo de que si las series de Taylor de $f(t)$ alrededor de $t = a$ es

$$f(t) = \sum_{n=0}^{\infty} c_n(t-a)^n$$

entonces

$$c_n = \frac{1}{n!}\frac{d^n}{dt^n}f(t)\bigg|_{t=a}$$

3.16. Pruebe el teorema 3-9, página 87.

Dado que X y Y son independientes, cualquier función de X y cualquier función de Y son independientes. Por lo tanto,

$$M_{X+Y}(t) = E[e^{t(X+Y)}] = E(e^{tX}e^{tY}) = E(e^{tX})E(e^{tY}) = M_X(t)M_Y(t)$$

3.17. La variable aleatoria X puede asumir los valores 1 y -1 cada uno con probabilidad de $\frac{1}{2}$. Encuentre (*a*) la función generadora de momento, (*b*) los primeros cuatro momentos alrededor del origen.

(*a*)
$$E(e^{tX}) = e^{t(1)}\left(\frac{1}{2}\right) + e^{t(-1)}\left(\frac{1}{2}\right) = \frac{1}{2}(e^t + e^{-t})$$

(*b*) tenemos
$$e^t = 1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} + \cdots$$

$$e^{-t} = 1 - t + \frac{t^2}{2!} - \frac{t^3}{3!} + \frac{t^4}{4!} - \cdots$$

Entonces (*1*) $$\frac{1}{2}(e^t + e^{-t}) = 1 + \frac{t^2}{2!} + \frac{t^4}{4!} + \cdots$$

Pero (*2*) $$M_X(t) = 1 + \mu t + \mu_2' \frac{t^2}{2!} + \mu_3' \frac{t^3}{3!} + \mu_4' \frac{t^4}{4!} + \cdots$$

Entonces, comparando (*1*) y (*2*), tenemos

$$\mu = 0, \qquad \mu_2' = 1, \qquad \mu_3' = 0, \qquad \mu_4' = 1, \ldots$$

Los momentos impares son todos cero, y los momentos pares son todos uno.

3.18. Una variable aleatoria X tiene función de densidad dada por

$$f(x) = \begin{cases} 2e^{-2x} & x \geqq 0 \\ 0 & x < 0 \end{cases}$$

Encuentre (*a*) la función generadora de momento, (*b*) los primeros cuatro momentos alrededor del origen.

(*a*) $$M(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx$$

$$= \int_0^{\infty} e^{tx}(2e^{-2x})\, dx = 2\int_0^{\infty} e^{(t-2)x}\, dx$$

$$= \frac{2e^{(t-2)x}}{t-2}\Bigg|_0^{\infty} = \frac{2}{2-t}, \quad \text{suponiendo } t < 2$$

(*b*) Si $|t| < 2$ tenemos

$$\frac{2}{2-t} = \frac{1}{1-t/2} = 1 + \frac{t}{2} + \frac{t^2}{4} + \frac{t^3}{8} + \frac{t^4}{16} + \cdots$$

Pero $$M(t) = 1 + \mu t + \mu_2' \frac{t^2}{2!} + \mu_3' \frac{t^3}{3!} + \mu_4' \frac{t^4}{4!} + \cdots$$

por consiguiente, al comparar términos, $\mu = \frac{1}{2}$, $\mu_2' = \frac{1}{2}$, $\mu_3' = \frac{3}{4}$, $\mu_4' = \frac{3}{2}$.

3.19. Encuentre los primeros cuatro momentos (*a*) alrededor del origen, (*b*) alrededor de la media, para una variable aleatoria X que tiene función de densidad

$$f(x) = \begin{cases} 4x(9 - x^2)/81 & 0 \leqq x \leqq 3 \\ 0 & \text{de otra manera} \end{cases}$$

(*a*) $$\mu_1' = E(X) = \frac{4}{81}\int_0^3 x^2(9 - x^2)\, dx = \frac{8}{5} = \mu$$

$$\mu_2' = E(X^2) = \frac{4}{81}\int_0^3 x^3(9 - x^2)\, dx = 3$$

$$\mu_3' = E(X^3) = \frac{4}{81}\int_0^3 x^4(9 - x^2)\, dx = \frac{216}{35}$$

$$\mu_4' = E(X^4) = \frac{4}{81}\int_0^3 x^5(9 - x^2)\, dx = \frac{27}{2}$$

(*b*) Usando el resultado (*27*), página 86, tenemos

$$\mu_1 = 0$$

$$\mu_2 = 3 - \left(\frac{8}{5}\right)^2 = \frac{11}{25} = \sigma^2$$

$$\mu_3 = \frac{216}{35} - 3(3)\left(\frac{8}{5}\right) + 2\left(\frac{8}{5}\right)^3 = -\frac{32}{875}$$

$$\mu_4 = \frac{27}{2} - 4\left(\frac{216}{35}\right)\left(\frac{8}{5}\right) + 6(3)\left(\frac{8}{5}\right)^2 - 3\left(\frac{8}{5}\right)^4 = \frac{3693}{8750}$$

FUNCIONES CARACTERÍSTICAS

3.20. Encuentre la función característica de la variable aleatoria X del problema 3.17.

La función característica está dada por

$$E(e^{i\omega X}) = e^{i\omega(1)}\left(\frac{1}{2}\right) + e^{i\omega(-1)}\left(\frac{1}{2}\right) = \frac{1}{2}\left(e^{i\omega} + e^{-i\omega}\right) = \cos\omega$$

usando la fórmula de Euler,

$$e^{i\theta} = \cos\theta + i\,\text{sen}\,\theta \qquad e^{-i\theta} = \cos\theta - i\,\text{sen}\,\theta$$

donde $\theta = \omega$. El resultado también puede obtenerse a partir del problema 3.17(*a*), remplazando $t = i\omega$.

3.21. Encuentre la función característica de la variable aleatoria X con función de densidad dada por

$$f(x) = \begin{cases} 1/2a & |x| < a \\ 0 & \text{de otra manera} \end{cases}$$

La función característica está dada por

$$E(e^{i\omega X}) = \int_{-\infty}^{\infty} e^{i\omega x} f(x)\,dx = \frac{1}{2a}\int_{-a}^{a} e^{i\omega x}\,dx$$

$$= \frac{1}{2a}\frac{e^{i\omega x}}{i\omega}\Bigg|_{-a}^{a} = \frac{e^{ia\omega} - e^{-ia\omega}}{2ia\omega} = \frac{\text{sen}\,a\omega}{a\omega}$$

usando las fórmulas de Euler (*véase* el problema 3.20) con $\theta = a\omega$.

3.22. Encuentre la función característica de la variable aleatoria X con función de densidad $f(x) = ce^{-a|x|}$ $-\infty < x < \infty$, donde $a > 0$, y c es una constante apropiada.

Dado que $f(x)$ es una función de densidad, debemos tener que

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

de manera que

$$c\int_{-\infty}^{\infty} e^{-a|x|}\,dx = c\left[\int_{-\infty}^{0} e^{-a(-x)}\,dx + \int_{0}^{\infty} e^{-a(x)}\,dx\right]$$

$$= c\frac{e^{ax}}{a}\Bigg|_{-\infty}^{0} + c\frac{e^{-ax}}{-a}\Bigg|_{0}^{\infty} = \frac{2c}{a} = 1$$

Entonces $c = a/2$. Por consiguiente, la función característica está dada por

$$E(e^{i\omega X}) = \int_{-\infty}^{\infty} e^{i\omega x} f(x)\,dx$$

$$= \frac{a}{2}\left[\int_{-\infty}^{0} e^{i\omega x}\, e^{-a(-x)}\,dx + \int_{0}^{\infty} e^{i\omega x} e^{-a(x)}\,dx\right]$$

$$= \frac{a}{2}\left[\int_{-\infty}^{0} e^{(a+i\omega)x}\,dx + \int_{0}^{\infty} e^{-(a-i\omega)x}\,dx\right]$$

$$= \frac{a}{2}\,\frac{e^{(a+i\omega)x}}{a+i\omega}\Bigg|_{-\infty}^{0} + a\,\frac{e^{-(a-i\omega)x}}{-(a-i\omega)}\Bigg|_{0}^{\infty}$$

$$= \frac{a}{2(a+i\omega)} + \frac{a}{2(a-i\omega)} = \frac{a^2}{a^2+\omega^2}$$

COVARIANZA Y COEFICIENTE DE CORRELACIÓN

3.23. Pruebe el teorema 3-14, página 89.

Por definición la covarianza de X y Y es

$$\sigma_{XY} = \mathrm{Cov}\,(X, Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

$$= E[XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y]$$

$$= E(XY) - \mu_X E(Y) - \mu_Y E(X) + E(\mu_X\mu_Y)$$

$$= E(XY) - \mu_X\mu_Y - \mu_Y\mu_X + \mu_X\mu_Y$$

$$= E(XY) - \mu_X\mu_Y$$

$$= E(XY) - E(X)E(Y)$$

3.24. Pruebe el teorema 3-15, página 89.

Si X y Y son independientes, entonces $E(XY) = E(X)E(Y)$. Por consiguiente, según el problema 3.23,

$$\sigma_{XY} = \mathrm{Cov}\,(X, Y) = E(XY) - E(X)E(Y) = 0$$

3.25. Encuentre (*a*) $E(X)$, (*b*) $E(Y)$, (*c*) $E(XY)$, (*d*) $E(X^2)$, (*e*) $E(Y^2)$, (*f*) Var (X), (*g*) Var (Y), (*h*) Cov (X, Y), (*i*) ρ, si las variables aleatorias X y Y se definen como en el problema 2.8, página 53.

(*a*)
$$E(X) = \sum_x \sum_y x f(x,y) = \sum_x x\left[\sum_y f(x,y)\right]$$

$$= (0)(6c) + (1)(14c) + (2)(22c) = 58c = \frac{58}{42} = \frac{29}{21}$$

(*b*)
$$E(Y) = \sum_x \sum_y y f(x,y) = \sum_y y\left[\sum_x f(x,y)\right]$$

$$= (0)(6c) + (1)(9c) + (2)(12c) + (3)(15c) = 78c = \frac{78}{42} = \frac{13}{7}$$

(*c*)
$$E(XY) = \sum_x \sum_y xy f(x,y)$$

$$= (0)(0)(0) + (0)(1)(c) + (0)(2)(2c) + (0)(3)(3c)$$
$$+ (1)(0)(2c) + (1)(1)(3c) + (1)(2)(4c) + (1)(3)(5c)$$
$$+ (2)(0)(4c) + (2)(1)(5c) + (2)(2)(6c) + (2)(3)(7c)$$

$$= 102c = \frac{102}{42} = \frac{17}{7}$$

$$(d) \qquad E(X^2) = \sum_x \sum_y x^2 f(x,y) = \sum_x x^2 \left[\sum_y f(x,y) \right]$$

$$= (0)^2(6c) + (1)^2(14c) + (2)^2(22c) = 102c = \frac{102}{42} = \frac{17}{7}$$

$$(e) \qquad E(Y^2) = \sum_x \sum_y y^2 f(x,y) = \sum_y y^2 \left[\sum_x f(x,y) \right]$$

$$= (0)^2(6c) + (1)^2(9c) + (2)^2(12c) + (3)^2(15c) = 192c = \frac{192}{42} = \frac{32}{7}$$

$$(f) \qquad \sigma_X^2 = \text{Var}\,(X) = E(X^2) - [E(X)]^2 = \frac{17}{7} - \left(\frac{29}{21}\right)^2 = \frac{230}{441}$$

$$(g) \qquad \sigma_Y^2 = \text{Var}\,(Y) = E(Y^2) - [E(Y)]^2 = \frac{32}{7} - \left(\frac{13}{7}\right)^2 = \frac{55}{49}$$

$$(h) \qquad \sigma_{XY} = \text{Cov}\,(X, Y) = E(XY) - E(X)E(Y) = \frac{17}{7} - \left(\frac{29}{21}\right)\left(\frac{13}{7}\right) = -\frac{20}{147}$$

$$(i) \qquad \rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{-20/147}{\sqrt{230/441}\,\sqrt{55/49}} = \frac{-20}{\sqrt{230}\,\sqrt{55}} = -0.2103 \text{ aproximadamente}$$

3.26. Resuelva el problema 3.25 si las variables aleatorias X y Y se definen como en el problema 2.33 página 67.

Usando $c = 1/210$, tenemos

$$(a) \qquad E(X) = \frac{1}{210} \int_{x=2}^{6} \int_{y=0}^{5} (x)(2x + y)\, dx\, dy = \frac{268}{63}$$

$$(b) \qquad E(Y) = \frac{1}{210} \int_{x=2}^{6} \int_{y=0}^{5} (y)(2x + y)\, dx\, dy = \frac{170}{63}$$

$$(c) \qquad E(XY) = \frac{1}{210} \int_{x=2}^{6} \int_{y=0}^{5} (xy)(2x + y)\, dx\, dy = \frac{80}{7}$$

$$(d) \qquad E(X^2) = \frac{1}{210} \int_{x=2}^{6} \int_{y=0}^{5} (x^2)(2x + y)\, dx\, dy = \frac{1220}{63}$$

$$(e) \qquad E(Y^2) = \frac{1}{210} \int_{x=2}^{6} \int_{y=0}^{5} (y^2)(2x + y)\, dx\, dy = \frac{1175}{126}$$

$$(f) \qquad \sigma_X^2 = \text{Var}\,(X) = E(X^2) - [E(X)]^2 = \frac{1220}{63} - \left(\frac{268}{63}\right)^2 = \frac{5036}{3969}$$

$$(g) \qquad \sigma_Y^2 = \text{Var}\,(Y) = E(Y^2) - [E(Y)]^2 = \frac{1175}{126} - \left(\frac{170}{63}\right)^2 = \frac{16{,}225}{7938}$$

$$(h) \qquad \sigma_{XY} = \text{Cov}\,(X, Y) = E(XY) - E(X)E(Y) = \frac{80}{7} - \left(\frac{268}{63}\right)\left(\frac{170}{63}\right) = -\frac{200}{3969}$$

$$(i) \qquad \rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{-200/3969}{\sqrt{5036/3969}\,\sqrt{16{,}225/7938}} = \frac{-200}{\sqrt{2518}\,\sqrt{16{,}225}} = -0.03129 \text{ aproximadamente}$$

ESPERANZA CONDICIONAL, VARIANZA Y MOMENTOS

3.27. Encuentre la esperanza condicional de Y dado $X = 2$ en el problema 2.8, página 53.

Como en el problema 2.27, páginas 63 y 64, la función de probabilidad condicional de Y, dado X, $= 2$ es

$$f(y \mid 2) = \frac{4+y}{22}$$

Entonces la esperanza condicional de Y, dado $X = 2$, es

$$E(Y \mid X = 2) = \sum_y y\left(\frac{4+y}{22}\right)$$

donde la sumatoria se toma sobre todos los valores de y que correspondan a $X = 2$. Esto está dado por

$$E(Y \mid X = 2) = (0)\left(\frac{4}{22}\right) + 1\left(\frac{5}{22}\right) + 2\left(\frac{6}{22}\right) + 3\left(\frac{7}{22}\right) = \frac{19}{11}$$

3.28. Encuentre la esperanza condicional de (a) Y dado X, (b) X dado Y, en el problema 2.29, página 64.

(a) $$E(Y \mid X = x) = \int_{-\infty}^{\infty} y f_2(y \mid x)\, dy = \int_0^x y\left(\frac{2y}{x^2}\right) dy = \frac{2x}{3}$$

(b) $$E(X \mid Y = y) = \int_{-\infty}^{\infty} x f_1(x \mid y)\, dx = \int_y^1 x\left(\frac{2x}{1-y^2}\right) dx$$

$$= \frac{2(1-y^3)}{3(1-y^2)} = \frac{2(1+y+y^2)}{3(1+y)}$$

3.29. Encuentre la varianza condicional de Y dado X del problema 2.29, página 64.

La varianza requerida (segundo momento alrededor de la media) está dada por

$$E[(Y-\mu_2)^2 \mid X = x] = \int_{-\infty}^{\infty} (y-\mu_2)^2 f_2(y \mid x)\, dy = \int_0^x \left(y - \frac{2x}{3}\right)^2 \left(\frac{2y}{x^2}\right) dy = \frac{x^2}{18}$$

donde hemos usado el hecho de que $\mu_2 = E(Y \mid X = x) = 2x/3$ del problema 3.28(a).

DESIGUALDAD DE CHEBYSHEV

3.30. Pruebe la desigualdad de Chebyshev.

Presentaremos la demostración para las variables aleatorias continuas. La demostración para las variables discretas es similar si las integrales se remplazan por sumatorias. Si $f(x)$ es la función de densidad de X, entonces

$$\sigma^2 = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$$

Dado que el integrando es no negativo, el valor de la integral sólo puede decrecer cuando el rango de integración se reduce. Por consiguiente,

$$\sigma^2 \geqq \int_{|x-\mu| \geqq \epsilon} (x-\mu)^2 f(x)\, dx \geqq \int_{|x-\mu| \geqq \epsilon} \epsilon^2 f(x)\, dx = \epsilon^2 \int_{|x-\mu| \geqq \epsilon} f(x)\, dx$$

Pero la última integral es igual a $P(|X-\mu| \geqq \epsilon)$. Entonces,

$$P(|X-\mu| \geqq \epsilon) \leqq \frac{\sigma^2}{\epsilon^2}$$

3.31. Para la variable aleatoria del problema 3.18, (a) encuentre $P(|X-\mu| > 1)$. (b) Use la desigualdad de Chebyshev para obtener un límite superior de $P(|X-\mu| > 1)$ y compárelo con el resultado en (a).

(*a*) A partir del problema 3.18, $\mu = 1/2$. Entonces

$$P(|X-\mu|<1) = P\left(\left|X-\frac{1}{2}\right|<1\right) = P\left(-\frac{1}{2}<X<\frac{3}{2}\right)$$
$$= \int_0^{3/2} 2e^{-2x}\,dx = 1-e^{-3}$$

Por consiguiente, $$P\left(\left|X-\frac{1}{2}\right| \geqq 1\right) = 1-(1-e^{-3}) = e^{-3} = 0.04979$$

(*b*) A partir del problema 3.18, $\sigma^2 = \mu_2' - \mu^2 = 1/4$. La desigualdad de Chebyshev con $\epsilon = 1$ da

$$P(|X-\mu| \geqq 1) \leqq \sigma^2 = 0.25$$

Comparando con (*a*), podemos ver que el límite proporcionado por la desigualdad de Chebyshev aquí es bastante rudimentario. En la práctica, la desigualdad de Chebyshev se usa para proveer estimados cuando es inconveniente o imposible obtener valores exactos.

LA LEY DE LOS GRANDES NÚMEROS

3.32. Pruebe la ley de los grandes números planteada en el teorema 3-19, página 91.

Tenemos $$E(X_1) = E(X_2) = \cdots = E(X_n) = \mu$$

$$\text{Var}\,(X_1) = \text{Var}\,(X_2) = \cdots = \text{Var}\,(X_n) = \sigma^2$$

Entonces $$E\left(\frac{S_n}{n}\right) = E\left(\frac{X_1+\cdots+X_n}{n}\right) = \frac{1}{n}\left[E(X_1)+\cdots+E(X_n)\right] = \frac{1}{n}(n\mu) = \mu$$

$$\text{Var}\,(S_n) = \text{Var}\,(X_1+\cdots+X_n) = \text{Var}\,(X_1)+\cdots+\text{Var}\,(X_n) = n\sigma^2$$

de manera que $$\text{Var}\left(\frac{S_n}{n}\right) = \frac{1}{n^2}\,\text{Var}\,(S_n) = \frac{\sigma^2}{n}$$

donde usamos el teorema 3-5 y una extensión del teorema 3-7.

Por consiguiente, según la desigualdad de Chebyshev con $X = S_n/n$, tenemos

$$P\left(\left|\frac{S_n}{n}-\mu\right| \geqq \epsilon\right) \leqq \frac{\sigma^2}{n\epsilon^2}$$

Tomando el límite como $n \to \infty$, se obtiene, como se requiere,

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n}-\mu\right| \geqq \epsilon\right) = 0$$

OTRAS MEDIDAS DE LA TENDENCIA CENTRAL

3.33. La función de densidad de una variable aleatoria continua X es

$$f(x) = \begin{cases} 4x(9-x^2)/81 & 0 \leqq x \leqq 3 \\ 0 & \text{de otra manera} \end{cases}$$

(*a*) Encuentre la moda. (*b*) Encuentre la mediana. (*c*) Compare moda, mediana y media.

(*a*) La moda se obtiene encontrando dónde tiene la densidad $f(x)$ un máximo relativo. El máximo relativo de $f(x)$ ocurre donde la derivada es cero, es decir,

$$\frac{d}{dx}\left[\frac{4x(9-x^2)}{81}\right] = \frac{36-12x^2}{81} = 0$$

Entonces $x = \sqrt{3} = 1.73$ aproximadamente, lo cual es la moda requerida. Observe que éste da el máximo, puesto que la segunda derivada, $-24x/81$, es negativa para $x = \sqrt{3}$.

(*b*) La mediana es aquel valor a para el cual $P(X \leqq a) = 1/2$. Ahora, para $0 < a < 3$,

$$P(X \leqq a) = \frac{4}{81}\int_0^a x(9 - x^2)\,dx = \frac{4}{81}\left(\frac{9a^2}{2} - \frac{a^4}{4}\right)$$

Igualando esto a 1/2, encontramos que

$$2a^4 - 36a^2 + 81 = 0$$

a partir de lo cual

$$a^2 = \frac{36 \pm \sqrt{(36)^2 - 4(2)(81)}}{2(2)} = \frac{36 \pm \sqrt{648}}{4} = 9 \pm \frac{9}{2}\sqrt{2}$$

Por consiguiente, la mediana requerida, la cual debe ubicarse entre 0 y 3, está dada por

$$a^2 = 9 - \frac{9}{2}\sqrt{2}$$

a partir de lo cual $a = 1.62$ aproximadamente.

(*c*)
$$E(X) = \frac{4}{81}\int_0^3 x^2(9 - x^2)\,dx = \frac{4}{81}\left(3x^3 - \frac{x^5}{5}\right)\Big|_0^3 = 1.60$$

lo cual es prácticamente igual a la mediana. La moda, la media y la mediana se muestran en la figura 3-6.

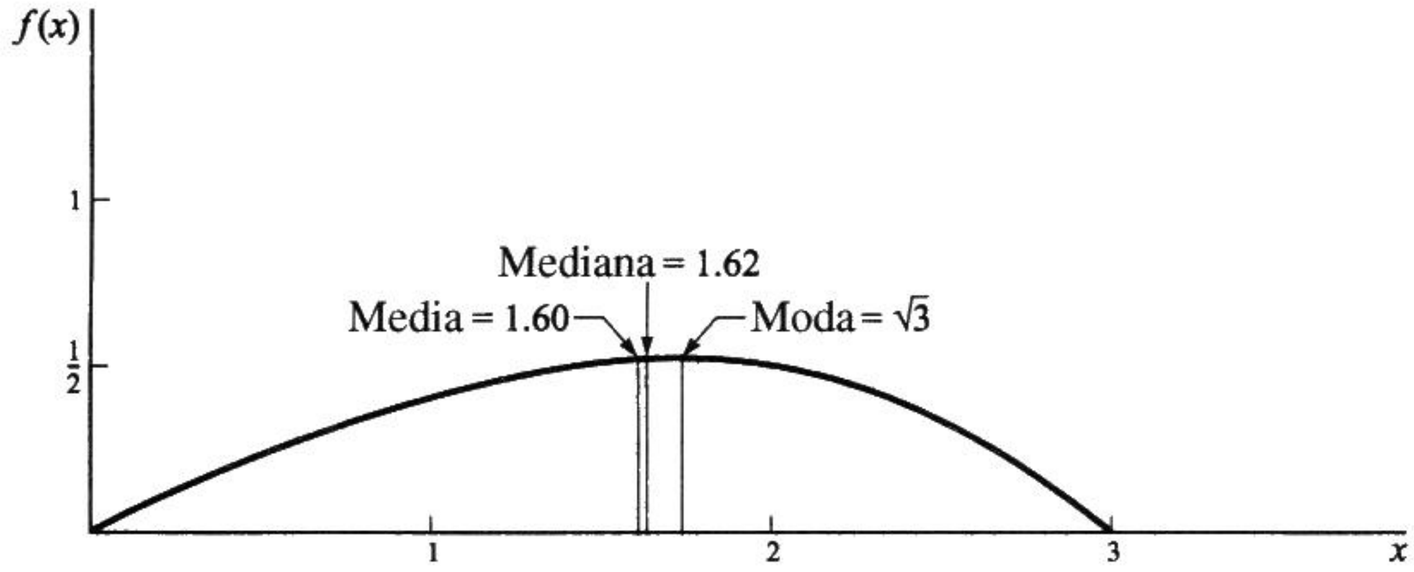


Fig. 3-6

3.34. Una variable aleatoria discreta tiene función de probabilidad $f(x) = 1/2^x$ donde $x = 1, 2, \ldots$. Encuentre (*a*) la moda, (*b*) la mediana, y (*c*) compárelas con la media.

(*a*) La moda es el valor de x que tiene asociada la mayor probabilidad. En este caso es $x = 1$, para lo cual la probabilidad es $1/2$.

(*b*) Si x es cualquier valor entre 1 y 2, $P(X < x) = \frac{1}{2}$ $P(X > x) = \frac{1}{2}$. Por consiguiente, *cualquier número* entre 1 y 2 puede representar la mediana. Por conveniencia, escogemos el punto central del intervalo, es decir, $3/2$.

(*c*) Tal como se encontró en el poblema 3.3, $\mu = 2$. Por consiguiente, el orden de las tres medidas es exactamente el inverso del orden del problema 3.33.

PERCENTILES

3.35. Determine los siguientes valores de percentil para el problema 3.33: (*a*) 10°, (*b*) 25-avo, (*c*) 75-avo.

A partir del problema 3.33(*b*) tenemos

$$P(X \leqq a) = \frac{4}{81}\left(\frac{9a^2}{2} - \frac{a^4}{4}\right) = \frac{18a^2 - a^4}{81}$$

(*a*) El 10º percentil es el valor a para el cual $P(X \leqq a) = 0.10$, es decir, la solución de $(18a^2 - x^4)/81 = 0.10$. Usando el método del problema 3.33, encontramos $a = 0.68$ aproximadamente.

(*b*) El 25-avo percentil es el valor a tal que $(18a^2 - x^4)/81 = 0.25$, y encontramos que $a = 1.098$ aproximadamente.

(*c*) El 75-avo percentil es el valor a tal que $(18a^2 - x^4)/81 = 0.75$, y encontramos que $a = 2.121$ aproximadamente.

OTRAS MEDIDAS DE DISPERSIÓN

3.36. Determine, (*a*) el recorrido semiintercuartílico, (*b*) la desviación media de la distribución del problema 3.33.

(*a*) Según el problema 3.33 el 25-avo y 75-avo valores de percentil son 1.098 y 2.121, respectivamente. Por lo tanto,

$$\text{recorrido semiintercuartílico} = = \frac{2.121 - 1.098}{2} = 0.51 \text{ aproximadamente}$$

(*b*) A partir de problema 3.33 la media es $\mu = 1.60 = 8/5$. Entonces

$$\text{Desviación media} = \text{D.M.} = E(|X - \mu|) = \int_{-\infty}^{\infty} |x - \mu| f(x)\, dx$$

$$= \int_0^3 \left| x - \frac{8}{5} \right| \left[\frac{4x}{81} (9 - x^2) \right] dx$$

$$= \int_0^{8/5} \left(\frac{8}{5} - x \right) \left[\frac{4x}{81} (9 - x^2) \right] dx + \int_{8/5}^3 \left(x - \frac{8}{5} \right) \left[\frac{4x}{81} (9 - x^2) \right] dx$$

$$= 0.555 \text{ aproximadamente}$$

SESGO Y CURTOSIS

3.37. Encuentre el coeficiente de (*a*) sesgo, (*b*) curtosis para la distribución del problema 3.19.

A partir del problema 3.19 (*b*) tenemos

$$\sigma^2 = \frac{11}{25} \qquad \mu_3 = -\frac{32}{875} \qquad \mu_4 = \frac{3693}{8750}$$

(*a*) Coeficiente de sesgo $= \alpha_3 = \dfrac{\mu_3}{\sigma^3} = -0.1253$

(*b*) Coeficiente de curtosis $= \alpha_4 = \dfrac{\mu_4}{\sigma^4} = 2.172$

Se deduce que hay un sesgo moderado a la izquierda, como lo indica la figura 3-6. También la distribución es de cierta manera menos puntuda que la distribución normal, la cual tiene una curtosis de 3.

PROBLEMAS DIVERSOS

3.38. Si $M(t)$ es la función generadora de momento para una variable aleatoria X, demuestre que la media es $\mu = M'(0)$ y la varianza es $\sigma^2 = M''(0) - [M'(0)]^2$.

A partir de (*32*), página 87, tenemos remplazando $r = 1$ y $r = 2$,

$$\mu_1' = M'(0) \qquad \mu_2' = M''(0)$$

Entonces a partir de (*27*)

$$\mu = M'(0) \qquad \mu_2 = \sigma^2 = M''(0) - [M'(0)]^2$$

3.39. Sea X una variable aleatoria que toma los valores $x_k = k$ con probabilidades p_k, donde $k = \pm 1, \ldots, \pm n$. (*a*) Encuentre la función característica $\phi(\omega)$ de X, (*b*) obtenga p_k en términos de $\phi(\omega)$.

(*a*) La función característica es

$$\phi(\omega) = E(e^{i\omega X}) = \sum_{k=-n}^{n} e^{i\omega x_k} p_k = \sum_{k=-n}^{n} p_k e^{ik\omega}$$

(*b*) Multiplique ambos lados de la expresión en (*a*) por $e^{-ij\omega}$ e integre con respecto a ω desde 0 hasta 2π. Entonces

$$\int_{\omega=0}^{2\pi} e^{-ij\omega}\phi(\omega)\,d\omega = \sum_{k=-n}^{n} p_k \int_{\omega=0}^{2\pi} e^{i(k-j)\omega}\,d\omega = 2\pi p_j$$

puesto que

$$\int_{\omega=0}^{2\pi} e^{i(k-j)\omega}\,d\omega = \begin{cases} \left.\dfrac{e^{i(k-j)\omega}}{i(k-j)}\right|_0^{2\pi} = 0 & k \neq j \\ 2\pi & k = j \end{cases}$$

Por consiguiente,

$$p_j = \frac{1}{2\pi}\int_{\omega=0}^{2\pi} e^{-ij\omega}\phi(\omega)\,d\omega$$

o, remplazando j por k,

$$p_k = \frac{1}{2\pi}\int_{\omega=0}^{2\pi} e^{-ik\omega}\phi(\omega)\,d\omega$$

Con frecuencia llamamos $\sum_{k=-n}^{n} p_k e^{ik\omega}$ (donde n puede ser teóricamente infinita) la serie de Fourier de $\phi(\omega)$ y p_k los coeficientes de Fourier. Para una variable aleatoria continua, la serie de Fourier se remplaza por la integral de Fourier (*véase* página 88).

3.40. Use el problema 3.39 para obtener la distribución de probabilidad de una variable aleatoria X cuya función característica es $\phi(\omega) = \cos\omega$.

A partir del problema 3.39

$$\begin{aligned} p_k &= \frac{1}{2\pi}\int_{\omega=0}^{2\pi} e^{-ik\omega}\cos\omega\,d\omega \\ &= \frac{1}{2\pi}\int_{\omega=0}^{2\pi} e^{-ik\omega}\left[\frac{e^{i\omega}+e^{-i\omega}}{2}\right]d\omega \\ &= \frac{1}{4\pi}\int_{\omega=0}^{2\pi} e^{i(1-k)\omega}\,d\omega + \frac{1}{4\pi}\int_{\omega=0}^{2\pi} e^{-i(1+k)\omega}\,d\omega \end{aligned}$$

Si $k = 1$, encontramos $p_1 = \frac{1}{2}$; si $k = -1$, encontramos $p_{-1} = \frac{1}{2}$. Para todos los demás valores de k, tenemos $p_k = 0$. Por consiguiente, la variable aleatoria está dada por

$$X = \begin{cases} 1 & \text{probabilidad } 1/2 \\ -1 & \text{probabilidad } 1/2 \end{cases}$$

Como verificación vea el problema 3.20.

3.41. Encuentre el coeficiente de (*a*) sesgo, (*b*) curtosis de la distribución definida por la *curva normal*, con densidad

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad -\infty < x < \infty$$

(*a*) La distribución tiene la apariencia de la figura 3-7. Por simetría, $\mu_1' = \mu = 0$ y $\mu_3' = 0$. Por consiguiente el coeficiente de sesgo es cero.

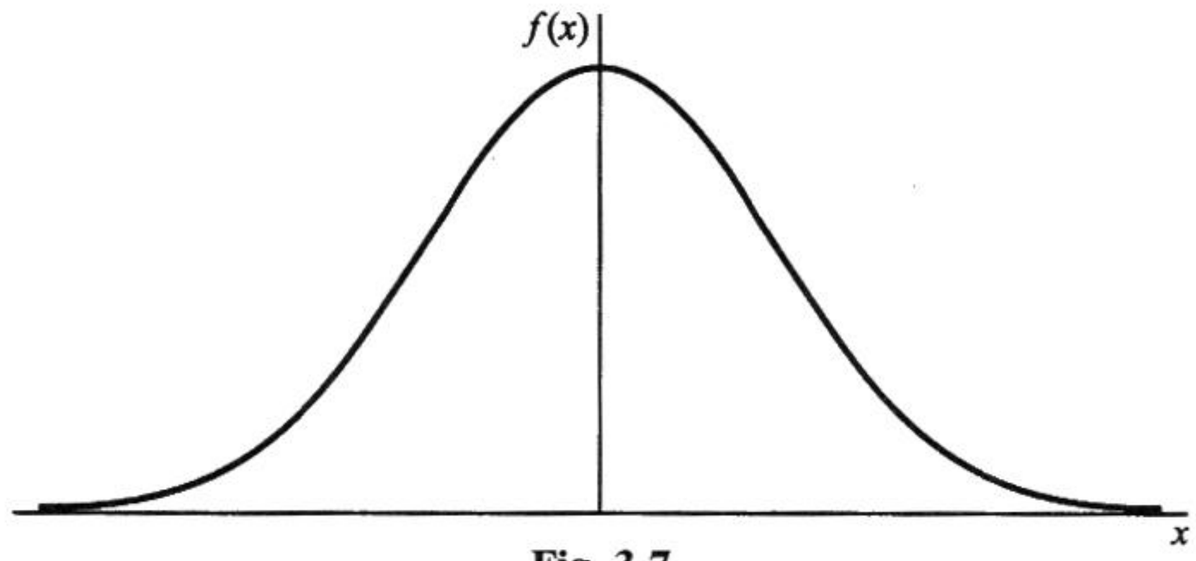


Fig. 3-7

(*b*) Tenemos

$$\mu_2' = E(X^2) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2 e^{-x^2/2}\,dx = \frac{2}{\sqrt{2\pi}}\int_0^{\infty} x^2 e^{-x^2/2}\,dx$$

$$= \frac{2}{\sqrt{\pi}}\int_0^{\infty} v^{1/2} e^{-v}\,dv$$

$$= \frac{2}{\sqrt{\pi}}\,\Gamma\left(\frac{3}{2}\right) = \frac{2}{\sqrt{\pi}}\cdot\frac{1}{2}\,\Gamma\left(\frac{1}{2}\right) = 1$$

donde hicimos la transformación $x^2/2 = v$ y usamos las propiedades de la función gamma dada en (*2*) y (*5*) del apéndice A. De igual manera obtenemos

$$\mu_4' = E(X^4) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^4 e^{-x^2/2}\,dx = \frac{2}{\sqrt{2\pi}}\int_0^{\infty} x^4 e^{-x^2/2}\,dx$$

$$= \frac{4}{\sqrt{\pi}}\int_0^{\infty} v^{3/2} e^{-v}\,dv$$

$$= \frac{4}{\sqrt{\pi}}\,\Gamma\left(\frac{5}{2}\right) = \frac{4}{\sqrt{\pi}}\cdot\frac{3}{2}\cdot\frac{1}{2}\Gamma\left(\frac{1}{2}\right) = 3$$

Ahora

$$\sigma^2 = E[(X-\mu)^2] = E(X^2) = \mu_2' = 1$$

$$\mu_4 = E[(X-\mu)^4] = E(X^4) = \mu_4' = 3$$

Así el coeficiente de la curtosis es

$$\frac{\mu_4}{\sigma^4} = 3$$

3.42. Pruebe que $-1 \leqq \rho \leqq 1$ (*véase* página 89).

Para cualquier constante real c, tenemos

$$E[\{Y - \mu_Y - c(X-\mu)\}^2] \geqq 0$$

Ahora, el lado izquierdo puede escribirse

$$E[(Y-\mu_Y)^2] + c^2E[(X-\mu_X)^2] - 2cE[(X-\mu_X)(Y-\mu_Y)] = \sigma_Y^2 + c^2\sigma_X^2 - 2c\sigma_{XY}$$

$$= \sigma_Y^2 + \sigma_X^2\left(c^2 - \frac{2c\sigma_{XY}}{\sigma_X^2}\right)$$

$$= \sigma_Y^2 + \sigma_X^2\left(c - \frac{\sigma_{XY}}{\sigma_X^2}\right)^2 - \frac{\sigma_{XY}^2}{\sigma_X^2}$$

$$= \frac{\sigma_X^2\sigma_Y^2 - \sigma_{XY}^2}{\sigma_X^2} + \sigma_X^2\left(c - \frac{\sigma_{XY}}{\sigma_X^2}\right)^2$$

Con el fin de que esta última cantidad sea mayor que o igual a cero para cada valor de c, debemos tener

$$\sigma_X^2\sigma_Y^2 - \sigma_{XY}^2 \geqq 0 \qquad \text{o} \qquad \frac{\sigma_{XY}^2}{\sigma_X^2\sigma_Y^2} \leqq 1$$

lo cual es equivalente a $\rho^2 \leqq 1$ ó $-1 \leqq \rho \leqq 1$.

Problemas suplementarios

ESPERANZA DE VARIABLES ALEATORIAS

3.43. Una variable aleatoria X se define por $X = \begin{cases} -2 & \text{probabilidad } 1/3 \\ 3 & \text{probabilidad } 1/2 \\ 1 & \text{probabilidad } 1/6 \end{cases}$. Encuentre (*a*) $E(X)$, (*b*) $E(2X + 5)$, (*c*) $E(X$

3.44. Sea X una variable aleatoria definida por la función de densidad $f(x) = \begin{cases} 3x^2 & 0 \leqq x \leqq 1 \\ 0 & \text{de otra manera} \end{cases}$.

Encuentre (*a*) $E(X)$, (*b*) $E(3X - 2)$, (*c*) $E(X^2)$.

3.45. La función de densidad de una variable aleatoria X es $f(x) = \begin{cases} e^{-x} & x \geqq 0 \\ 0 & \text{de otra manera} \end{cases}$.

Encuentre (*a*) $E(X)$, (*b*) $E(X^2)$, (*c*) $E[(X - 1)^2]$.

3.46. ¿Cuál es el número esperado de puntos que aparecerá en 3 lanzamientos sucesivos de un dado balanceado? ¿Parece razonable su respuesta? Explique.

3.47. Una variable aleatoria X tiene la función de densidad $f(x) = \begin{cases} e^{-x} & x \geqq 0 \\ 0 & x < 0 \end{cases}$. Encuentre $E(e^{2X/3})$.

3.48. Sean X y Y variables aleatorias independientes cada una con función de densidad

$$f(u) = \begin{cases} 2e^{-2u} & u \geqq 0 \\ 0 & \text{de otra manera} \end{cases}$$

Encuentre (*a*) $E(X + Y)$, (*b*) $E(X^2 + Y^2)$, (*c*) $E(XY)$.

3.49. En el problema 3.48, ¿son (*a*) $E(X + Y) = E(X) + E(Y)$, (*b*) $E(XY) = E(X)E(Y)$? Explique.

3.50. Sean X y Y variables aleatorias con función de densidad conjunta

$$f(x,y) = \begin{cases} \frac{3}{5}x(x+y) & 0 \leqq x \leqq 1,\ 0 \leqq y \leqq 2 \\ 0 & \text{de otra manera} \end{cases}$$

Encuentre (*a*) $E(X)$, (*b*) $E(Y)$, (*c*) $E(X + Y)$, (*d*) $E(XY)$.

3.51. En el problema 3.50 , ¿son (*a*) $E(X + Y) = E(X) + E(Y)$, (*b*) $E(XY) = E(X)E(Y)$? Explique.

3.52. Sean X y Y variables aleatorias con función de densidad conjunta

$$f(x,y) = \begin{cases} 4xy & 0 \leqq x \leqq 1,\ 0 \leqq y \leqq 1 \\ 0 & \text{de otra manera} \end{cases}$$

Encuentre (*a*) $E(X)$, (*b*) $E(Y)$, (*c*) $E(X + Y)$, (*d*) $E(XY)$.

3.53. En el problema 3.52, ¿son (*a*) $E(X + Y) = E(X) + E(Y)$, (*b*) $E(XY) = E(X)E(Y)$? Explique.

3.54. Sea $f(x,y) = \begin{cases} \frac{1}{4}(2x+y) & 0 \leqq x \leqq 1,\ 0 \leqq y \leqq 2 \\ 0 & \text{de otra manera} \end{cases}$. Encuentre (a) $E(X)$, (b) $E(Y)$, (c) $E(X^2)$, (d) $E(Y^2)$, (e) $E(X+Y)$, (f) $E(XY)$.

3.55. Sean X y Y variables aleatorias independientes tales que

$$X = \begin{cases} 1 & \text{prob. } 1/3 \\ 0 & \text{prob. } 2/3 \end{cases} \quad ; \quad Y = \begin{cases} 2 & \text{prob. } 3/4 \\ -3 & \text{prob. } 1/4 \end{cases}$$

Encuentre (a) $E(3X+2Y)$, (b) $E(2X^2 - Y^2)$, (c) $E(XY)$, (d) $E(X^2Y)$.

3.56. Sean $X_1, X_2, \ldots, X_n$ n variables aleatorias distribuidas idénticamente, tal que

$$X_k = \begin{cases} 1 & \text{prob. } 1/2 \\ 2 & \text{prob. } 1/3 \\ -1 & \text{prob. } 1/6 \end{cases}$$

Encuentre (a) $E(X_1 + X_2 + \cdots + X_n)$, (b) $E(X_1^2 + X_2^2 + \ldots + X_n^2)$.

VARIANZA Y DESVIACIÓN ESTÁNDAR

3.57. Encuentre (a) la varianza, (b) la desviación estándar del número de puntos que aparecen en un lanzamiento sencillo de un dado balanceado.

3.58. Sea X la variable aleatoria con función de densidad

$$f(x) = \begin{cases} 1/4 & -2 \leqq x \leqq 2 \\ 0 & \text{de otra manera} \end{cases}$$

Encuentre (a) Var (X), (b) σ_X.

3.59. Sea X la variable aleatoria con función de densidad

$$f(x) = \begin{cases} e^{-x} & x \geqq 0 \\ 0 & \text{de otra manera} \end{cases}$$

Encuentre (a) Var (X), (b) σ_X.

3.60. Encuentre la varianza y la desviación estándar para la variable aleatoria X de (a) problema 3.43, (b) problema 3.44.

3.61. Una variable aleatoria X tiene $E(X) = 2$, $E(X^2) = 8$. Encuentre (a) Var (X), (b) σ_X.

3.62. Si una variable aleatoria X es tal que $E[(X-1)^2] = 10$, $E[(X-2)^2] = 6$, encuentre (a) $E(X)$ (b) Var (X), (c) σ_X.

MOMENTOS Y FUNCIONES GENERADORAS DE MOMENTOS

3.63. Encuentre (a) la función generadora de momento de la variable aleatoria

$$X = \begin{cases} 1/2 & \text{prob. } 1/2 \\ -1/2 & \text{prob. } 1/2 \end{cases}$$

y (b) los primeros cuatro momentos alrededor del origen.

3.64. (a) Encuentre la función generadora de momento de una variable aleatoria X con función de densidad

$$f(x) = \begin{cases} x/2 & 0 \leqq x \leqq 2 \\ 0 & \text{de otra manera} \end{cases}$$

(b) Use la función generadora de momento de (a) para encontrar los primeros cuatro momentos alrededor del origen.

3.65. Encuentre los primeros cuatro momentos alrededor de la media en (*a*) problema 3.43, (*b*) problema 3.44.

3.66. (*a*) Encuentre la función generadora de momento de una variable aleatoria con función de densidad

$$f(x) = \begin{cases} e^{-x} & x \geqq 0 \\ 0 & \text{de otra manera} \end{cases}$$

y (*b*) determine los primeros cuatro momentos alrededor del origen.

3.67. Encuentre en el problema 3.66 los primeros cuatro momentos alrededor de la media.

3.68. Sea X con función de densidad $f(x) = \begin{cases} 1/(b-a) & a \leqq x \leqq b \\ 0 & \text{de otra manera} \end{cases}$. Encuentre el k-ésimo momento alrededor de (*a*) el origen, (*b*) la media.

3.69. Si $M(t)$ es la función generadora de momento de la variable aleatoria X, pruebe que el tercero y cuarto momentos alrededor de la media están dados por

$$\mu_3 = M'''(0) - 3M''(0)M'(0) + 2[M'(0)]^3$$

$$\mu_4 = M^{(\text{iv})}(0) - 4M'''(0)M'(0) + 6M''(0)[M'(0)]^2 - 3[M'(0)]^4$$

FUNCIONES CARACTERÍSTICAS

3.70. Encuentre la función característica de la variable aleatoria $X = \begin{cases} a & \text{prob. } p \\ b & \text{prob. } q = 1-p \end{cases}$.

3.71. Encuentre la función característica de una variable aleatoria X que tiene función de densidad

$$f(x) = \begin{cases} 1/2a & |x| \leqq a \\ 0 & \text{de otra manera} \end{cases}$$

3.72. Encuentre la función característica de una variable aleatoria con función de densidad

$$f(x) = \begin{cases} x/2 & 0 \leqq x \leqq 2 \\ 0 & \text{de otra manera} \end{cases}$$

3.73. Sea $X_k = \begin{cases} 1 & \text{prob. } 1/2 \\ -1 & \text{prob. } 1/2 \end{cases}$ variables aleatorias independientes ($k = 1, 2, \ldots, n$). Pruebe que la función característica de la variable aleatoria

$$\frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}}$$

es $[\cos(\omega/\sqrt{n})]^n$.

3.74. Pruebe que a medida que $n \to \infty$ la función característica del problema 3.73 se acerca a $e^{-\omega^2/2}$. (Ayuda: Use el logaritmo de la función característica y utilice la regla de L'Hospital.)

COVARIANZA Y COEFICIENTE DE CORRELACIÓN

3.75. Sean X y Y variables aleatorias con función de densidad conjunta

$$f(x,y) = \begin{cases} x+y & 0 \leqq x \leqq 1,\ 0 \leqq y \leqq 1 \\ 0 & \text{de otra manera} \end{cases}$$

Encuentre (*a*) Var (X), (*b*) Var (Y), (*c*) σ_X, (*d*) σ_Y, (*e*) σ_{XY}, (*f*) ρ.

3.76. Desarrolle el problema 3.75 si la función de densidad conjunta es $f(x,y) = \begin{cases} e^{-(x+y)} & x \geqq 0,\ y \geqq 0 \\ 0 & \text{de otra manera} \end{cases}$.

3.77. Encuentre (*a*) Var (*X*), (*b*) Var (*Y*), (*c*) σ_X, (*d*) σ_Y, (*e*) σ_{XY}, (*f*) ρ, para las variables aleatorias del problema 2.56.

3.78. Desarrolle el problema 3.77 para las variables aleatorias del problema 2.94.

3.79. Encuentre (*a*) la covarianza, (*b*) el coeficiente de correlación de dos variables aleatorias X y Y si $E(X) = 2$, $E(Y) = 3$, $E(XY) = 10$, $E(X^2) = 9$, $E(Y^2) = 16$.

3.80. El coeficiente de correlación de dos variables aleatorias X y Y es $-\frac{1}{4}$, mientras que sus varianzas son 3 y 5. Encuentre la covarianza.

ESPERANZA CONDICIONAL, VARIANZA Y MOMENTOS

3.81. La función de densidad conjunta de X y Y es

$$f(x,y) = \begin{cases} x+y & 0 \leqq x \leqq 1,\ 0 \leqq y \leqq 1 \\ 0 & \text{de otra manera} \end{cases}$$

Encuentre la esperanza condicional de (*a*) Y dado X, (*b*) X dado Y.

3.82. Desarrolle el problema 3.81 si $f(x,y) = \begin{cases} 2e^{-(x+2y)} & x \geqq 0,\ y \geqq 0 \\ 0 & \text{de otra manera} \end{cases}$

3.83. La función de probabilidad conjunta de X y Y está dada en la tabla 2-9, página 77. Encuentre la esperanza condicional de (*a*) Y dado X, (*b*) X dado Y.

3.84. Encuentre la varianza condicional de (*a*) Y dado X, (*b*) X dado Y, para la distribución del problema 3.81.

3.85. Desarrolle el problema 3.84 para la distribución del problema 3.82.

3.86. Desarrolle el problema 3.84 para la distribución del problema 2.94.

DESIGUALDAD DE CHEBYSHEV

3.87. Una variable aleatoria X tiene media 3 y varianza 2. Use la desigualdad de Chebyshev para obtener un límite superior para (*a*) $P(|X-3| \geqq 2)$, (*b*) $P(|X-3| \geqq 1)$.

3.88. Pruebe la desigualdad de Chebyshev para una variable discreta X. (Ayuda: *Véase* el problema 3.30)

3.89. Una variable aleatoria X tiene función de densidad $f(x) = \frac{1}{2}e^{-|x|}$, $-\infty < x < \infty$. (*a*) Encuentre $P(|X-\mu| > 2)$. (*b*) Use la desigualdad de Chebyshev para obtener un límite superior en $P(|X-\mu| > 2)$ y compárelo con el resultado en (*a*).

LEY DE LOS GRANDES NÚMEROS

3.90. Pruebe que la ley (débil) de los grandes números puede establecerse como

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| < \epsilon\right) = 1$$

e interprete.

3.91. Sean X_k $(k = 1, \ldots, n)$ variables aleatorias independientes tales que

$$X_k = \begin{cases} 1 & \text{prob. } p \\ 0 & \text{prob. } q = 1 - p \end{cases}$$

(*a*) Si interpretamos x_k como el número de caras en el k-ésimo lanzamiento de una moneda, ¿cuál interpretación puede darse a $S_n = X_1 + \cdots + X_n$?

(*b*) Pruebe que en este caso la ley de los grandes números se reduce a

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - p\right| \geqq \epsilon\right) = 0$$

e interprete el resultado.

OTRAS MEDIDAS DE LA TENDENCIA CENTRAL

3.92. Encuentre (*a*) la moda, (*b*) la mediana de una variable aleatoria X con función de densidad

$$f(x) = \begin{cases} e^{-x} & x \geqq 0 \\ 0 & \text{de otra manera} \end{cases}$$

y (*c*) compare con la media.

3.93. Desarrolle el problema 3.92 si la función de densidad es

$$f(x) = \begin{cases} 4x(1 - x^2) & 0 \leqq x \leqq 1 \\ 0 & \text{de otra manera} \end{cases}$$

3.94. Encuentre (*a*) la mediana, (*b*) la moda de una variable aleatoria X definida como

$$X = \begin{cases} 2 & \text{prob. } 1/3 \\ -1 & \text{prob. } 2/3 \end{cases}$$

y (*c*) compare con la media.

3.95. Encuentre (*a*) la mediana, (*b*) la moda de un conjunto de números 1, 3, 2, 1, 5, 6, 3, 3 y (*c*) compare con la media.

PERCENTILES

3.96. Encuentre el (*a*) 25-avo, (*b*) 75-avo valor de percentil para la variable aleatoria con función de densidad

$$f(x) = \begin{cases} 2(1 - x) & 0 \leqq x \leqq 1 \\ 0 & \text{de otra manera} \end{cases}$$

3.97. Encuentre el (*a*) 10°, (*b*) 25-avo, (*c*) 75-avo, (*d*) 90-avo valor de percentil para la variable aleatoria con función de densidad

$$f(x) = \begin{cases} c(x - x^3) & 0 < x < 1 \\ 0 & \text{de otra manera} \end{cases}$$

donde c es una constante apropiada.

OTRAS MEDIDAS DE DISPERSIÓN

3.98. Encuentre (*a*) el recorrido semiintercuartílico, (*b*) la desviación media para la variable aleatoria del problema 3.96.

3.99. Desarrolle el problema 3.98 para la variable aleatoria del problema 3.97.

3.100. Encuentre la desviación media de la variable aleatoria X en cada uno de los casos siguientes.

$$(a)\ \ f(x) = \begin{cases} e^{-x} & x \geq 0 \\ 0 & \text{en otro caso} \end{cases} \qquad (b)\ \ f(x) = \frac{1}{\pi(1+x^2)}, \qquad -\infty < x < \infty.$$

3.101. Obtenga la probabilidad de que la variable aleatoria X difiera de su media por más que el recorrido semiintercuartílico en el caso de (a) problema 3.96, (b) problema 3.100(a).

SESGO Y CURTOSIS

3.102. Encuentre el coeficiente de (a) sesgo, (b) curtosis para la distribución del problema 3.100(a).

3.103. Si

$$f(x) = \begin{cases} c\left(1 - \frac{|x|}{a}\right) & |x| \leqq a \\ 0 & |x| > a \end{cases}$$

donde c es una constante apropiada, es la función de densidad de X, encuentre el coeficiente de (a) sesgo, (b) curtosis.

3.104. Encuentre el coeficiente de (a) sesgo, (b) curtosis para la distribución con función de densidad

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geqq 0 \\ 0 & x < 0 \end{cases}$$

PROBLEMAS DIVERSOS

3.105. Sea X una variable aleatoria que puede tomar los valores 2, 1 y 3 con probabilidades respectivas de 1/3, 1/6 y 1/2. Encuentre (a) la media, (b) la varianza, (c) la función generadora de momento, (d) la función característica, (e) el tercer momento alrededor de la media.

3.106. Desarrolle el problema 3.105 si X tiene función de densidad

$$f(x) = \begin{cases} c(1-x) & 0 < x < 1 \\ 0 & \text{de otra manera} \end{cases}$$

donde c es una constante apropiada.

3.107. Se lanzan sucesivamente tres dados, que se suponen balanceados. Encuentre (a) la media, (b) la varianza de la suma.

3.108. Sea X una variable aleatoria con función de densidad

$$f(x) = \begin{cases} cx & 0 \leqq x \leqq 2 \\ 0 & \text{de otra manera} \end{cases}$$

donde c es una constante apropiada. Encuentre (a) la media, (b) la varianza, (c) la función generadora de momento, (d) la función característica, (e) el coeficiente de sesgo, (f) el coeficiente de curtosis.

3.109. La función de densidad conjunta de X y Y está dada por

$$f(x,y) = \begin{cases} cxy & 0 < x < 1,\ 0 < y < 1 \\ 0 & \text{de otra manera} \end{cases}$$

Encuentre (a) $E(X^2 + Y^2)$, (b) $E(\sqrt{X^2 + Y^2})$.

3.110. Desarrolle el problema 3.109 si X y Y son variables aleatorias idénticamente distribuidas con función de densidad $f(u) = (2\pi)^{-1/2} e^{-u^2/2}$, $-\infty < u < \infty$.

3.111. Sea X una variable aleatoria con función de densidad

$$f(x) = \begin{cases} \frac{1}{2} & -1 < x < 1 \\ 0 & \text{de otra manera} \end{cases}$$

y $Y = X^2$. Encuentre (*a*) $E(X)$, (*b*) $E(Y)$, (*c*) $E(XY)$.

Respuestas a los problemas suplementarios

3.43. (*a*) 1 (*b*) 7 (*c*) 6

3.44. (*a*) 3/4 (*b*) 1/4 (*c*) 3/5

3.45. (*a*) 1 (*b*) 2 (*c*) 1

3.46. 10.5

3.47. 3

3.48. (*a*) 1 (*b*) 1 (*c*) 1/4

3.50. (*a*) 7/10 (*b*) 6/5 (*c*) 19/10 (*d*) 5/6

3.52. (*a*) 2/3 (*b*) 2/3 (*c*) 4/3 (*d*) 4/9

3.54. (*a*) 7/12 (*b*) 7/6
(*c*) 5/12 (*d*) 5/3 (*e*) 7/4 (*f*) 2/3

3.55. (*a*) 5/2 (*b*) −55/12 (*c*) 1/4 (*d*) 1/4

3.56. (*a*) n (*b*) $2n$

3.57. (*a*) 35/12 (*b*) $\sqrt{35/12}$

3.58. (*a*) 4/3 (*b*) $\sqrt{4/3}$

3.59. (*a*) 1 (*b*) 1

3.60. (*a*) $\text{Var}(X) = 5$, $\sigma_X = \sqrt{5}$ (*b*) $\text{Var}(X) = 3/80$, $\sigma_X = \sqrt{15}/20$

3.61. (*a*) 4 (*b*) 2

3.62. (*a*) 7/2 (*b*) 15/4 (*c*) $\sqrt{15}/2$

3.63. (*a*) $\frac{1}{2}(e^{t/2} + e^{-t/2}) = \cosh(t/2)$ (*b*) $\mu = 0$, $\mu_2' = 1$, $\mu_3' = 0$, $\mu_4' = 1$

3.64. (*a*) $(1 + 2te^{2t} - e^{2t})/2t^2$ (*b*) $\mu = 4/3$, $\mu_2' = 2$, $\mu_3' = 16/5$, $\mu_4' = 16/3$

3.65. (*a*) $\mu_1 = 0$, $\mu_2 = 5$, $\mu_3 = -5$, $\mu_4 = 35$
(*b*) $\mu_1 = 0$, $\mu_2 = 3/80$, $\mu_3 = -121/160$, $\mu_4 = 2307/8960$

3.66. (*a*) $1/(1 - t)$, $|t| < 1$ (*b*) $\mu = 1$, $\mu_2' = 2$, $\mu_3' = 6$, $\mu_4' = 24$

3.67. $\mu_1 = 0$, $\mu_2 = 1$, $\mu_3 = 2$, $\mu_4 = 33$

3.68. (*a*) $(b^{k+1} - a^{k+1})/(k + 1)(b - a)$ (*b*) $[1 + (-1)^k](b - a)^k/2^{k+1}(k + 1)$

3.70. $pe^{i\omega a} + qe^{i\omega b}$

3.71. $(\text{sen } a\omega)/a\omega$

3.72. $(e^{2i\omega} - 2i\omega e^{2i\omega} - 1)/2\omega^2$

3.75. (*a*) 11/144 (*b*) 11/144 (*c*) $\sqrt{11}/12$ (*d*) $\sqrt{11}/12$ (*e*) −1/144 (*f*) −1/11

3.76. (*a*) 1 (*b*) 1 (*c*) 1 (*d*) 1 (*e*) 0 (*f*) 0

3.77. (*a*) 73/960 (*b*) 73/960 (*c*) $\sqrt{73/960}$ (*d*) $\sqrt{73/960}$ (*e*) $-1/64$ (*f*) $-15/73$

3.78. (*a*) 233/324 (*b*) 233/324 (*c*) $\sqrt{233}/18$ (*d*) $\sqrt{233}/18$ (*e*) $-91/324$ (*f*) $-91/233$

3.79. (*a*) 4 (*b*) $4/\sqrt{35}$

3.80. $-\sqrt{15}/4$

3.81. (*a*) $(3x+2)/(6x+3)$ para $0 \leqq x \leqq 1$ (*b*) $(3y+2)/(6y+3)$ para $0 \leqq y \leqq 1$

3.82. (*a*) 1/2 para $x \geqq 0$ (*b*) 1 para $y \geqq 0$

3.83. (*a*)

X	0	1	2
$E(Y \mid X)$	4/3	1	5/7

(*b*)

Y	0	1	2
$E(X \mid Y)$	4/3	7/6	1/2

3.84. (*a*) $\dfrac{6x^2+6x+1}{18(2x+1)^2}$ para $0 \leqq x \leqq 1$ (*b*) $\dfrac{6y^2+6y+1}{18(2y+1)^2}$ para $0 \leqq y \leqq 1$

3.85. (*a*) 1/9 (*b*) 1

3.86. (*a*)

X	0	1	2
$\mathrm{Var}(Y \mid X)$	5/9	4/5	24/49

(*b*)

Y	0	1	2
$\mathrm{Var}(X \mid Y)$	5/9	29/36	7/12

3.87. (*a*) 1/2 (*b*) 2 (inútil)

3.89. (*a*) e^{-2} (*b*) 0.5

3.92. (*a*) $+0$ (*b*) $\ln 2$ (*c*) 1

3.93. (*a*) $1/\sqrt{3}$ (*b*) $\sqrt{1-(1/\sqrt{2})}$ (*c*) 8/15

3.94. (*a*) no existe (*b*) -1 (*c*) 0

3.95. (*a*) 3 (*b*) 3 (*c*) 3

3.96. (*a*) $1-\frac{1}{2}\sqrt{3}$ (*b*) 1/2

3.97. (*a*) $\sqrt{1-(3/\sqrt{10})}$ (*b*) $\sqrt{1-(\sqrt{3}/2)}$ (*c*) $\sqrt{1/2}$ (*d*) $\sqrt{1-(1/\sqrt{10})}$

3.98. (*a*) 1 (*b*) $(\sqrt{3}-1)/4$ (*c*) 16/81

3.99. (*a*) 1 (*b*) 0.17 (*c*) 0.051

3.100. (*a*) $1-2e^{-1}$ (*b*) no existe

3.101. (*a*) $(5-2\sqrt{3})/3$ (*b*) $(3-2e^{-1}\sqrt{3})/3$

3.102. (*a*) 2 (*b*) 9

3.103. (*a*) 0 (*b*) $24/5a$

3.104. (*a*) 2 (*b*) 9

3.105. (*a*) 7/3 (*b*) 5/9 (*c*) $(e^t+2e^{2t}+3e^{3t})/6$ (*d*) $(e^{i\omega}+2e^{2i\omega}+3e^{3i\omega})/6$ (*e*) $-7/27$

3.106. (*a*) 1/3 (*b*) 1/18 (*c*) $2(e^t - 1 - t)/t^2$ (*d*) $-2(e^{i\omega} - 1 - i\omega)/\omega^2$ (*e*) 1/135

3.107. (*a*) 21/2 (*b*) 35/4

3.108. (*a*) 4/3 (*c*) $(1 + 2te^{2t} - e^{2t})/2t^2$ (*e*) $-2\sqrt{18}/15$
(*b*) 2/9 (*d*) $-(1 + 2i\omega e^{2i\omega} - e^{2i\omega})/2\omega^2$ (*f*) 12/5

3.109. (*a*) 1 (*b*) $8(2\sqrt{2} - 1)/15$

3.110. (*a*) 2 (*b*) $\sqrt{2\pi}/2$

3.111. (*a*) 0 (*b*) 1/3 (*c*) 0

CAPÍTULO 4

Distribuciones de probabilidad especial

LA DISTRIBUCIÓN BINOMIAL

Supongamos que estamos realizando un experimento como lanzar repetidamente una moneda o un dado o sacar repetidamente una canica de una urna. Cada lanzamiento o selección se llama una *prueba*. En cualquier prueba sencilla habrá una probabilidad asociada con un evento en particular, como cara en una moneda, 4 en un dado o sacar una canica roja. En algunos casos esta probabilidad no cambiará de una prueba a la siguiente (como ocurre con el lanzamiento de una moneda o de un dado). Se dice que tales pruebas son *independientes* y se denominan *pruebas de Bernoulli*, por James Bernoulli quien las investigó a finales del siglo XVII.

Sea p la probabilidad de que ocurra un evento en un ensayo sencillo de Bernoulli (llamado *probabilidad de éxito*). Entonces, $q = 1 - p$ es la probabilidad de que tal evento no ocurra en una prueba sencilla (llamada *probabilidad de fracaso*). La probabilidad de que el evento ocurra exactamente x veces en n pruebas (es decir, éxitos y ocurran $n - x$ fracasos), está dada por la función de probabilidad

$$f(x) = P(X = x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x} \tag{1}$$

donde la variable aleatoria X denota el número de éxitos en n pruebas y $x = 0, 1, ..., n$.

EJEMPLO 4.1. La probabilidad de obtener exactamente 2 caras en 6 lanzamientos de una moneda balanceada es

$$P(X = 2) = \binom{6}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{6-2} = \frac{6!}{2!4!}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{6-2} = \frac{15}{64}$$

La función de probabilidad discreta (1) con frecuencia se denomina distribución binomial dado que para $x = 0, 1, 2, ..., n$, le corresponden términos sucesivos en la *expansión binomial*.

$$(q+p)^n = q^n + \binom{n}{1}q^{n-1}p + \binom{n}{2}q^{n-2}p^2 + \cdots + p^n = \sum_{x=0}^{n}\binom{n}{x}p^x q^{n-z} \tag{2}$$

El caso especial de una distribución binomial con $n = 1$ también recibe el nombre de *distribución de Bernoulli*.

ALGUNAS PROPIEDADES DE LA DISTRIBUCIÓN BINOMIAL

Algunas de las propiedades importantes de la distribución binomial se presentan en la tabla 4-1.

Tabla 4-1

Media	$\mu = np$
Varianza	$\sigma^2 = npq$
Desviación estándar	$\sigma = \sqrt{npq}$
Coeficiente de sesgo	$\alpha_3 = \dfrac{q-p}{\sqrt{npq}}$
Coeficiente de curtosis	$\alpha_4 = 3 + \dfrac{1-6pq}{npq}$
Función generadora de momento	$M(t) = (q + pe^t)^n$
Función característica	$\phi(\omega) = (q + pe^{i\omega})^n$

EJEMPLO 4.2. En 100 lanzamientos de una moneda balanceada, el número esperado o número medio de caras es $\mu = (100)(\frac{1}{2}) = 50$, mientras que la desviación estándar es $\sigma = \sqrt{(100)(\frac{1}{2})(\frac{1}{2})} = 5$.

LA LEY DE LOS GRANDES NÚMEROS PARA LAS PRUEBAS DE BERNOULLI

La ley de los grandes números, páginas 90-91, tiene una interpretación interesante en el caso de las pruebas de Bernoulli y se presenta en el teorema siguiente.

Teorema 4-1 **(Ley de los grandes números para pruebas de Bernoulli):** Sea X la variable aleatoria que indica el número de éxitos en n pruebas de Bernoulli, de manera que X/n es la proporción de éxitos. Entonces, si p es la probabilidad de éxito y ϵ es cualquier número positivo,

$$\lim_{n\to\infty} P\left(\left|\frac{X}{n} - p\right| \geqq \epsilon\right) = 0 \tag{3}$$

En otras palabras, a largo plazo se vuelve extremadamente probable que la proporción de éxito, X/n, sea tan cercana como usted quiera a la probabilidad de éxito de una prueba sencilla, p. En cierto sentido esta ley justifica el uso de la definición empírica de probabilidad de la página 5. La *ley fuerte de los grandes números* (página 90) da un resultado más determinante, el cual establece que con probabilidad de uno, $\lim_{n\to\infty} X/n = p$, es decir, X/n en realidad *converge a p* excepto en un número despreciable de casos.

LA DISTRIBUCIÓN NORMAL

Uno de los ejemplos más importantes de una distribución de probabilidad continua es la *distribución normal*, llamada a veces la *distribución gaussiana*. La función de densidad para esta distribución está dada por

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < \infty \tag{4}$$

donde μ y σ son la media y la desviación estándar, respectivamente. La función de distribución correspondiente está dada por

$$F(x) = P(X \leqq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(v-\mu)^2/2\sigma^2}\, dv \tag{5}$$

Si X tiene la función de distribución dada en (5), decimos que la variable aleatoria X está distribuida normalmente con media μ y varianza σ^2.

Sea Z la variable estandarizada correspondiente a X, es decir, si tenemos

$$Z = \frac{X - \mu}{\sigma} \tag{6}$$

entonces la media o valor esperado de Z es 0 y la varianza es 1. En tales casos la función de densidad para Z se puede obtener de (4) remplazando formalmente $\mu = 0$ y $\sigma = 1$, se obtiene

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \tag{7}$$

Esto con frecuencia se refiere a la *función de densidad normal estándar*. La función de distribución correspondiente está dada por

$$F(z) = P(Z \leqq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2}\, du = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_{0}^{x} e^{-u^2/2}\, du \tag{8}$$

Algunas veces llamamos al valor z de la variable estandarizada Z el *valor estándar*. La función $F(z)$ se relaciona con la función de error tabulada, fer(z). Tenemos

$$\text{fer}\,(z) = \frac{2}{\sqrt{\pi}} \int_{0}^{z} e^{-u^2}\, du \qquad \text{y} \qquad F(z) = \frac{1}{2}\left[1 + \text{fer}\left(\frac{z}{\sqrt{2}}\right)\right] \tag{9}$$

En la figura 4-1 se muestra la gráfica de la función de densidad (7), que algunas veces recibe el nombre de *curva normal estándar*. En esta gráfica indicamos las áreas dentro de 1, 2 y 3 desviaciones estándar de la media (o sea, entre $z = -1$ y $+1$; $z = -2$ y $+2$; $z = -3$ y $+3$), que equivalen a 68.27%, 95.45% y 99.73%, del área total, respectivamente, siendo ésta última igual a uno. Esto significa que

$$P(-1 \leqq Z \leqq 1) = 0.6827, \qquad P(-2 \leqq Z \leqq 2) = 0.9545, \qquad P(-3 \leqq Z \leqq 3) = 0.9973 \tag{10}$$

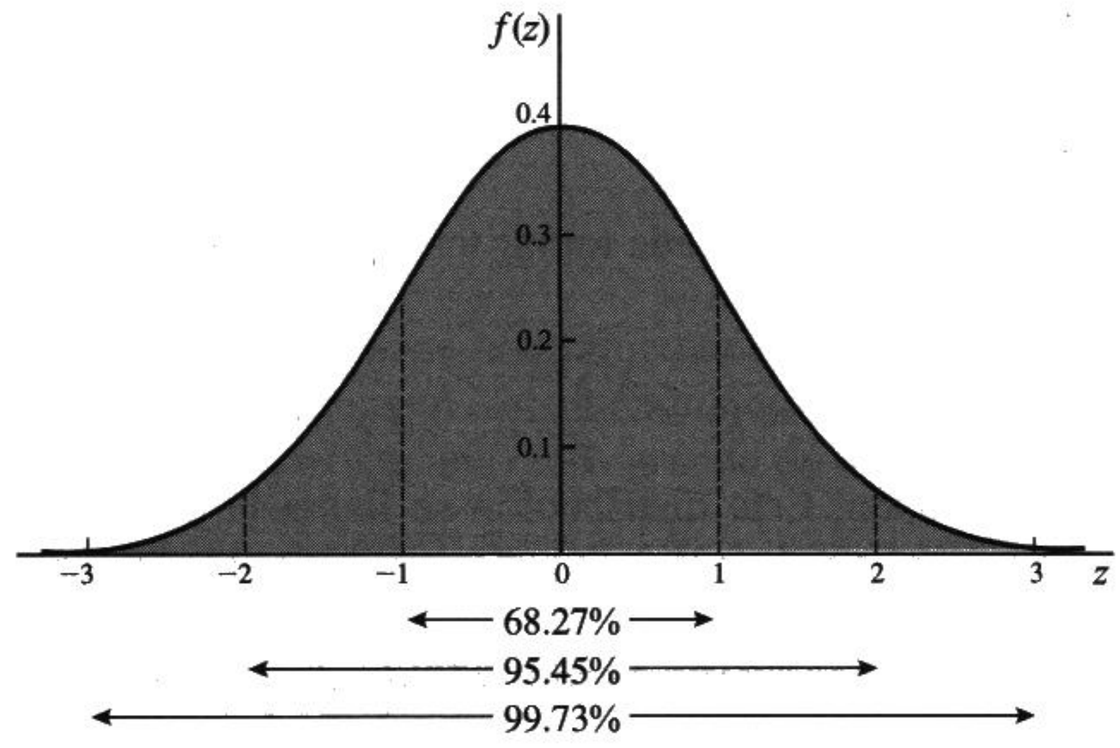


Fig. 4-1

En el apéndice C hay una tabla en la que encontrará las áreas bajo esta curva limitada por las ordenadas $z = 0$ y cualquier valor positivo de z. A partir de esta tabla las áreas entre dos ordenadas cualesquiera se pueden encontrar usando la simetría de la curva alrededor de $z = 0$.

ALGUNAS PROPIEDADES DE LA DISTRIBUCIÓN NORMAL

En la tabla 4-2 se listan algunas propiedades importantes de la distribución normal general.

Tabla 4-2

Media	μ
Varianza	σ^2
Desviación estándar	σ
Coeficiente de sesgo	$\alpha_3 = 0$
Coeficiente de curtosis	$\alpha_4 = 3$
Función generadora de momento	$M(t) = e^{\mu t + (\sigma^2 t^2/2)}$
Función característica	$\phi(\omega) = e^{i\mu\omega - (\sigma^2\omega^2/2)}$

RELACIÓN ENTRE LAS DISTRIBUCIONES BINOMIAL Y NORMAL

Si n es grande y p ni q están muy cerca de cero, la distribución binomial puede aproximarse bastante a una distribución normal con una variable aleatoria estandarizada dada por

$$Z = \frac{X - np}{\sqrt{npq}} \tag{11}$$

Aquí X es la variable aleatoria que da el número de éxitos en las pruebas de Bernoulli y p es la probabilidad de éxito. La aproximación mejora con el incremento de n y es exacta en el caso limitante (*véase* problema 4.17.) En la práctica, la aproximación es muy buena si tanto np como nq son mayores que 5. El hecho de que la distribución binomial se aproxime a la distribución normal puede expresarse escribiendo

$$\lim_{n\to\infty} P\left(a \leqq \frac{X - np}{\sqrt{npq}} \leqq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-u^2/2}\, du \tag{12}$$

En palabras, decimos que la variable aleatoria estandarizada $(X - np)/\sqrt{npq}$ es asintóticamente normal.

LA DISTRIBUCIÓN DE POISSON

Sea X una variable aleatoria discreta que puede tomar los valores 0, 1, 2, ... tal que la función de probabilidad de X está dada por

$$f(x) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad x = 0, 1, 2, \ldots \tag{13}$$

donde λ es una constante positiva dada. Esta distribución se llama la *distribución de Poisson* (por S. D. Poisson, quien la descubrió a comienzos del siglo XIX), y la variable aleatoria que tiene esta distribución se dice que *está distribuida según Poisson*.

Los valores de $f(x)$ en (*13*) pueden obtenerse usando el apéndice G, el cual da los valores de $e^{-\lambda}$ para diferentes valores de λ.

ALGUNAS PROPIEDADES DE LA DISTRIBUCIÓN DE POISSON

En la tabla 4-3 se listan algunas propiedades importantes de la distribución de Poisson.

RELACIÓN ENTRE LAS DISTRIBUCIONES POISSON Y BINOMIAL

En la distribución binomial (*1*), si n es grande mientras la propiedad de ocurrencia de un evento p es cercana a cero, de manera que $q = 1 - p$ es cercano a 1, el evento recibe el nombre de *evento raro*. En la

Tabla 4-3

Media	$\mu = \lambda$
Varianza	$\sigma^2 = \lambda$
Desviación estándar	$\sigma = \sqrt{\lambda}$
Coeficiente de sesgo	$\alpha_3 = 1/\sqrt{\lambda}$
Coeficiente de curtosis	$\alpha_4 = 3 + (1/\lambda)$
Función generadora de momento	$M(t) = e^{\lambda(e^t - 1)}$
Función característica	$\phi(\omega) = e^{\lambda(e^{j\omega} - 1)}$

práctica debemos considerar un evento como raro si el número de pruebas es al menos 50 ($n \geqq 50$), mientras que np es menor que 5. Para tales casos la distribución binomial se aproxima bastante a la distribución de Poisson (*13*) con $\lambda = np$. Esto es de esperarse si comparamos las tablas 4-1 y 4-3, dado que remplazando $\lambda = np$, $q \approx 1$ y $p \approx 0$ en la tabla 4-1, obtenemos los resultados de la tabla 4-3.

RELACIÓN ENTRE LAS DISTRIBUCIONES NORMAL Y POISSON

Dado que existe una relación entre las distribuciones binomial y normal y entre la binomial y la de Poisson, podríamos esperar que hubiera también una relación entre la distribución de Poisson y la distribución normal. De hecho, éste es el caso. Podemos demostrar que si X es la variable aleatoria de Poisson de (*13*) y $(X - \lambda)/\sqrt{\lambda}$ es la variable aleatoria estandarizada correspondiente, entonces

$$\lim_{\lambda \to \infty} P\left(a \leqq \frac{X - \lambda}{\sqrt{\lambda}} \leqq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-u^2/2}\, du \tag{14}$$

es decir, la distribución de Poisson se aproxima a la distribución normal a medida que $\lambda \to \infty$ o $(X - \lambda)/\sqrt{\lambda}$ es *asintóticamente normal*.

EL TEOREMA DEL LÍMITE CENTRAL

La similitud entre (*12*) y (*14*) nos conduce naturalmente a preguntarnos si existen otras distribuciones, además de la binomial y de Poisson, que tengan distribución normal como caso límite. El siguiente y extraordinario teorema revela que en realidad existe una clase grande de distribuciones que tienen esta propiedad.

Teorema 4-2 **(Teorema del límite central):** Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes que están idénticamente distribuidas (es decir, todas con la *misma* función de probabilidad en el caso discreto o función de densidad en el caso continuo) y tienen media μ y varianza σ^2 finitas. Entonces, si $S_n = X_1, + X_2 + \cdots + X_n$ $(n = 1, 2, \ldots)$,

$$\lim_{n \to \infty} P\left(a \leqq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leqq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-u^2/2}\, du \tag{15}$$

esto es, la variable aleatoria $(S_n - n\mu)/\sigma\sqrt{n}$, la cual es la correspondiente variable estandarizada de S_n, es asintóticamente normal.

El teorema también es verdadero bajo condiciones más generales; por ejemplo, se cumple cuando $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes con la misma media y la misma varianza, pero no necesariamente idénticamente distribuidas.

LA DISTRIBUCIÓN MULTINOMIAL

Supongamos que los eventos $A_1, A_2, \ldots, A_k$ son mutuamente excluyentes y pueden ocurrir con las probabilidades respectivas $p_1, p_2, \ldots, p_k$, donde $p_1 + p_2 + \cdots + p_k = 1$. Si $X_1, X_2, \ldots, X_k$ son variables aleatorias que indican respectivamente el número de veces que $A_1, A_2, \ldots, A_k$ ocurren en un número n de pruebas, de manera que $X_1 + X_2 + \cdots + X_k = n$, entonces

$$P(X_1 = n_1, X_2 = n_2, \ldots, X_k = n_k) = \frac{n}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k} \tag{16}$$

donde $n_1 + n_2 + \cdots + n_k = n$, es la función de probabilidad conjunta de las variables aleatorias $X_1, \ldots, X_k$.

Esta distribución, la cual es una generalización de la distribución binomial, se llama *distribución multinomial*, puesto que (*16*) es el término general en la expansión multinomial de $(p_1 + p_2 + \cdots + p_k)^n$.

EJEMPLO 4.3. Si un dado balanceado se lanza 12 veces, la probabilidad de obtener exactamente dos veces 1, 2, 3, 4, 5 y 6 puntos es

$$P(X_1 = 2, X_2 = 2, \ldots, X_6 = 2) = \frac{12!}{2!2!2!2!2!2!} \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^2 = \frac{1925}{559{,}872} = 0.00344$$

El número esperado de veces que $A_1, A_2, \ldots, A_k$ ocurren en n pruebas son $np_1, np_2, \ldots, np_k$, respectivamente, es decir,

$$E(X_1) = np_1, \qquad E(X_2) = np_2, \qquad \ldots, \qquad E(X_k) = np_k \tag{17}$$

LA DISTRIBUCIÓN HIPERGEOMÉTRICA

Supongamos que una caja contiene b canicas blancas y r canicas rojas. Llevemos a cabo n pruebas de un experimento en el cual sacamos una canica al azar, observamos su color, y luego la devolvemos a la caja. Este tipo de experimento se denomina *muestreo con remplazo*. En tal caso, si X es la variable aleatoria que indica el número de canicas blancas que se sacan (éxitos) en n pruebas, entonces usando la distribución binomial (*1*) vemos que la probabilidad de exactamente n éxitos es

$$P(X = x) = \binom{n}{x} \frac{b^x r^{n-x}}{(b+r)^n}, \qquad x = 0, 1, \ldots, n \tag{18}$$

puesto que $p = b/(b + r)$, $q = 1 - p = r/(b + r)$.

Si modificamos lo anterior de manera que el muestreo se haga *sin remplazo*, es decir, las canicas no se devuelven a la caja después de haberlas sacado, entonces

$$P(X = x) = \frac{\binom{b}{x}\binom{r}{n-x}}{\binom{b+r}{n}}, \qquad x = \begin{matrix}\text{máx}\,(0, n-r), \ldots, \\ \text{mín}\,(n, b)\end{matrix} \tag{19}$$

Ésta es la *distribución hipergeométrica*. La media y la varianza para esta distribución son

$$\mu = \frac{nb}{b+r}, \qquad \sigma^2 = \frac{nbr(b+r-n)}{(b+r)^2(b+r-1)} \tag{20}$$

Si decimos que el número total de canicas blancas y rojas es N, mientras que las proporciones de canicas blancas y rojas son p y $q = 1 - p$, respectivamente, entonces

$$p = \frac{b}{b+r} = \frac{b}{N}, \qquad q = \frac{r}{b+r} = \frac{r}{N} \qquad \text{o} \qquad b = Np, \qquad r = Nq \tag{21}$$

de manera que (*19*) y (*20*) se convierten, respectivamente, en

$$P(X = x) = \frac{\binom{Np}{x}\binom{Nq}{n-x}}{\binom{N}{n}} \tag{22}$$

$$\mu = np, \qquad \sigma^2 = \frac{npq(N-n)}{N-1} \tag{23}$$

Observe que a medida que $N \to \infty$ (o N es grande en comparación con n), (*22*) se reduce a (*18*), lo cual puede escribirse

$$P(X = x) = \binom{n}{x} p^x q^{n-x} \tag{24}$$

y (*23*) se reduce a

$$\mu = np, \qquad \sigma^2 = npq \tag{25}$$

de acuerdo con las dos primeras entradas de la tabla 4-1, página 118. Los resultados son tal como podríamos esperar, ya que cuando N es grande, muestrear sin remplazo es prácticamente lo mismo que hacerlo con remplazo.

LA DISTRIBUCIÓN UNIFORME

Se dice que una variable aleatoria X está *distribuida uniformemente* en $a \leqq x \leqq b$ si su función de densidad es

$$f(x) = \begin{cases} 1/(b-a) & a \leqq x \leqq b \\ 0 & \text{de otra manera} \end{cases} \tag{26}$$

y la distribución se llama *distribución uniforme*.

La función de distribución está dada por

$$F(x) = P(X \leqq x) = \begin{cases} 0 & x < a \\ (x-a)/(b-a) & a \leqq x < b \\ 1 & x \geqq b \end{cases} \tag{27}$$

La media y la varianza son, respectivamente,

$$\mu = \frac{1}{2}(a+b), \qquad \sigma^2 = \frac{1}{12}(b-a)^2 \tag{28}$$

LA DISTRIBUCIÓN DE CAUCHY

Se dice que una variable aleatoria X tiene la *distribución de Cauchy* si la función de densidad de X es

$$f(x) = \frac{a}{\pi(x^2 + a^2)} \qquad a > 0, -\infty < x < \infty \tag{29}$$

Esta función de densidad es simétrica alrededor de $x = 0$, de manera que su mediana es cero. Sin embargo, no existen la media, la varianza ni los momentos superiores. Asimismo, la función generadora de momento no existe. Sin embargo, la función característica existe, y está dada por

$$\phi(\omega) = e^{-a\omega} \tag{30}$$

LA DISTRIBUCIÓN GAMMA

Se dice que una variable aleatoria X tiene la distribución gamma, o está distribuida gamma, si la función de densidad es

$$f(x) = \begin{cases} \dfrac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)} & x > 0 \\ 0 & x \leqq 0 \end{cases} \qquad (\alpha, \beta > 0) \tag{31}$$

donde $\Gamma(\alpha)$ es la *función gamma* (*véase* apéndice A). La media y la varianza están dadas por

$$\mu = \alpha\beta, \qquad \sigma^2 = \alpha\beta^2 \tag{32}$$

La función generadora de momento y la función característica están dadas, respectivamente, por

$$M(t) = (1 - \beta t)^{-\alpha}, \qquad \phi(\omega) = (1 - \beta i\omega)^{-\alpha} \tag{33}$$

LA DISTRIBUCIÓN BETA

Se dice que una variable aleatoria X tiene *distribución beta*, o está distribuida beta, si la función de densidad es

$$f(x) = \begin{cases} \dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} & 0 < x < 1 \\ 0 & \text{de otra manera} \end{cases} \qquad (\alpha, \beta > 0) \tag{34}$$

donde $B(\alpha, \beta)$ es la *función beta* (*véase* apéndice A). En vista de la relación (*9*), apéndice *A*, entre las funciones gamma y beta, la distribución beta también puede definirse por la función de densidad

$$f(x) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1} & 0 < x < 1 \\ 0 & \text{de otra manera} \end{cases} \tag{35}$$

donde α y β son positivas. La media y la varianza son

$$\mu = \frac{\alpha}{\alpha+\beta}, \qquad \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{36}$$

Para $\alpha > 1$, $\beta > 1$ hay una moda única en el valor

$$x_{\text{moda}} = \frac{\alpha - 1}{\alpha + \beta - 2} \tag{37}$$

LA DISTRIBUCIÓN CHI CUADRADO

Sean $X_1, X_2, \ldots, X_\nu$ ν variables aleatorias independientes distribuidas normalmente con media cero y varianza 1. Consideremos la variable aleatoria

$$\chi^2 = X_1^2 + X_2^2 + \cdots + X_\nu^2 \tag{38}$$

donde χ^2 se llama *chi cuadrado*. Entonces, podemos demostrar que para $x \geqq 0$,

$$P(\chi^2 \leqq x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \int_0^x u^{(\nu/2)-1} e^{-u/2}\, du \tag{39}$$

y $P(\chi^2 \leqq x) = 0$ para $x > 0$.

La distribución definida por (*39*) se llama *distribución chi cuadrado* y v es el *número de grados de libertad*. La distribución definida por (*39*) tiene una función de densidad correspondiente dada por

$$f(x) = \begin{cases} \dfrac{1}{2^{\nu/2}\Gamma(\nu/2)}\, x^{(\nu/2)-1} e^{-x/2} & x > 0 \\ 0 & x \leqq 0 \end{cases} \tag{40}$$

Se ha visto que la distribución chi cuadrado es un caso especial de la distribución gamma con $\alpha = \nu/2$ y $\beta = 2$. Por consiguiente,

$$\mu = \nu, \qquad \sigma^2 = 2\nu, \qquad M(t) = (1-2t)^{-\nu/2}, \qquad \phi(\omega) = (1-2i\omega)^{-\nu/2} \tag{41}$$

Para v grandes ($v \geqq 30$), podemos demostrar que $\sqrt{2\chi^2} - \sqrt{2\nu - 1}$ casi está distribuida normalmente con media 0 y varianza 1.

Los siguientes tres teoremas serán de gran utilidad en el estudio posterior:

Teorema 4-3: Sean $X_1, X_2, \ldots, X_\nu$ variables aleatorias independientes distribuidas normalmente con media 0 y varianza 1. Entonces $\chi^2 = X_1^2 + X_2^2 + \cdots + X_\nu^2$ tiene distribución chi cuadrado con v grados de libertad.

Teorema 4-4: Sean $U_1, U_2, \ldots, U_k$ variables aleatorias independientes que tienen distribución chi cuadrado con $v_1, v_2, \ldots, v_k$ grados de libertad, respectivamente. Entonces su suma $W = U_1 + U_2 + \cdots + U_k$ tiene distribución chi cuadrado con $v_1 + v_2 + \cdots + v_k$ grados de libertad.

Teorema 4-5: Sean V_1 y V_2 variables aleatorias independientes. Supongamos que V_1 tiene distribución chi cuadrado con v_1 grados de libertad, mientras que $V = V_1 + V_2$ tiene distribución chi cuadrado con v grados de libertad, donde $v > v_1$. Entonces V_2 tiene distribución chi cuadrado con $v - v_1$ grados de libertad.

En relación con la distribución chi cuadrado, la distribución t (adelante), la distribución F (página 126), y otras, es común en el estudio estadístico usar el *mismo símbolo* para la variable aleatoria y un valor de tal variable aleatoria. Por lo tanto, valores percentiles de la distribución chi cuadrado para v grados de libertad se denotan por $\chi^2_{p,\nu}$, o χ^2_p, o brevemente χ^2_p si v se sobreentiende, y no por $\chi_{p,\nu}$ o χ_p (*véase* apéndice E). Ésta es una notación ambigua, y el lector debe tener cuidado con ella, especialmente cuando cambie variables en las funciones de densidad.

LA DISTRIBUCIÓN *t* DE STUDENT

Si una variable aleatoria tiene función de densidad

$$f(t) = \frac{\Gamma\left(\dfrac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\dfrac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2} \qquad -\infty < t < \infty \tag{42}$$

se dice que tiene *distribución t de Student*, o *distribución t*, con v grados de libertad. Si v grande ($v \geqq 30$), la gráfica de $f(t)$ se aproxima muy de cerca a la curva normal estándar como se indica en la figura 4-2. Los valores percentiles de la distribución t para v grados de libertad se denotan por $t_{p,\nu}$ o t_p si v se sobreentiende. Vea el apéndice D en donde encontrará una tabla con dichos valores. Dado que la distribución t es simétrica, $t_{1-p} = -t_p$; por ejemplo, $t_{0.5} = -t_{0.95}$.

Para la distribución t tenemos

$$\mu = 0 \qquad \text{y} \qquad \sigma^2 = \frac{\nu}{\nu - 2} \qquad (\nu > 2). \tag{43}$$

El siguiente teorema es importante para el estudio futuro.

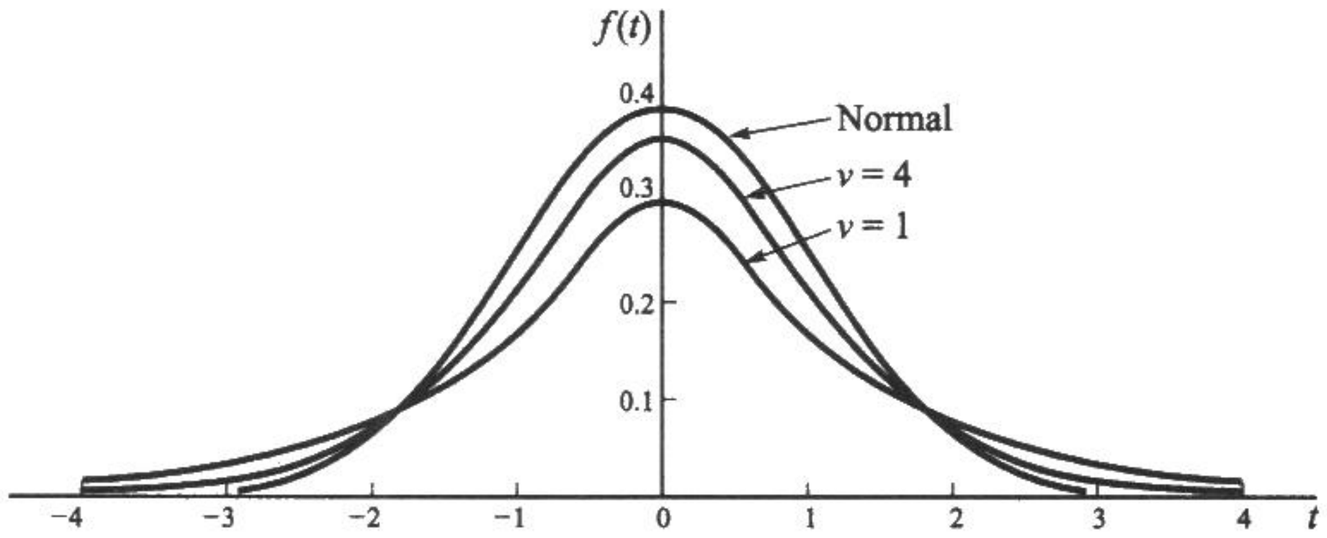


Fig. 4-2

Teorema 4-6: Sean Y y Z variables aleatorias independientes, donde Y está distribuida normalmente con media 0 y varianza 1, mientras que Z tiene distribución chi cuadrado con ν grados de libertad. Entonces la variable aleatoria

$$T = \frac{Y}{\sqrt{Z/\nu}} \tag{44}$$

tiene la distribución t con ν grados de libertad.

LA DISTRIBUCIÓN *F*

Se dice que una variable aleatoria tiene *distribución F* (por R. A. Fisher) *con* ν_1 *y* ν_2 *grados de libertad* si su función de densidad está dada por

$$f(u) = \begin{cases} \dfrac{\Gamma\left(\dfrac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\dfrac{\nu_1}{2}\right)\Gamma\left(\dfrac{\nu_2}{2}\right)} \nu_1^{\nu_1/2}\nu_2^{\nu_2/2}u^{(\nu_1/2)-1}(\nu_2+\nu_1 u)^{-(\nu_1+\nu_2)/2} & u > 0 \\ 0 & u \leqq 0 \end{cases} \tag{45}$$

Los valores percentiles de la distribución F con ν_1, ν_2 grados de libertad se denotan por F_{p,ν_1,ν_2}, o F_p si ν_1 y ν_2 se sobreentienden. En el apéndice F encontrará una tabla con dichos valores en los casos donde $p = 0.95$ y $p = 0.99$.

La media y la varianza están dadas, respectivamente, por

$$\mu = \frac{\nu_2}{\nu_2 - 2} \quad (\nu_2 > 2) \qquad \text{y} \qquad \sigma^2 = \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-4)(\nu_2-2)^2} \quad (\nu_2 > 4) \tag{46}$$

La distribución tiene una moda única en el valor

$$u_{\text{moda}} = \left(\frac{\nu_1 - 2}{\nu_1}\right)\left(\frac{\nu_2}{\nu_2+2}\right) \qquad (\nu_1 > 2) \tag{47}$$

Los siguientes teoremas son importantes en el trabajo futuro.

Teorema 4-7: Sean V_1 y V_2 variables aleatorias independientes que tienen distribución chi cuadrado con ν_1 y ν_2 grados de libertad, respectivamente. Entonces la variable aleatoria

$$V = \frac{V_1/\nu_1}{V_2/\nu_2} \tag{48}$$

tiene distribución F con ν_1 y ν_2 grados de libertad.

Teorema 4-8:

$$F_{1-p,\nu_2,\nu_1} = \frac{1}{F_{p,\nu_1,\nu_2}}$$

RELACIÓN ENTRE LAS DISTRIBUCIONES CHI CUADRADO, t Y F

Teorema 4-9: $$F_{1-p,1,\nu} = t^2_{1-(p/2),\nu}$$

Teorema 4-10: $$F_{p,\nu,\infty} = \frac{\chi^2_{p,\nu}}{\nu}$$

LA DISTRIBUCIÓN NORMAL BIVARIADA

Una generalización de la distribución normal a dos variables aleatorias continuas X y Y está dada por la función de densidad conjunta

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right] \Big/ 2(1-\rho^2)\right\} \tag{49}$$

donde $-\infty < x < \infty$, $-\infty < y < \infty$; μ_1, μ_2 son las medias de X y Y; σ_1, σ_2 son las desviaciones estándar de X y Y; y ρ es el coeficiente de correlación entre X y Y. Con frecuencia nos referimos a (*49*) como la *distribución normal bivariada*.

En cualquier distribución conjunta es necesaria la condición $\rho = 0$ para que haya independencia de las variables aleatorias (*véase* teorema 3-15). En el caso (*49*) esta condición también es suficiente (*véase* problema 4.51).

DIVERSAS DISTRIBUCIONES

En las distribuciones enumeradas a continuación, las constantes α, β, a, b, ... se toman como positivas a menos que se indique lo contrario. La función característica $\phi(\omega)$ se obtiene de la función generadora de momento, cuando está dada, remplazando $t = i\omega$.

1. DISTRIBUCIÓN GEOMÉTRICA.

$$f(x) = P(X = x) = pq^{x-1} \qquad x = 1, 2, \ldots$$

$$\mu = \frac{1}{p} \qquad \sigma^2 = \frac{q}{p^2} \qquad M(t) = \frac{pe^t}{1 - qe^t}$$

La variable aleatoria X representa el número de pruebas de Bernoulli hasta que ocurra el primer éxito, incluyendo la prueba en que ocurre éste. Aquí p es la probabilidad de éxito en una prueba sencilla.

2. DISTRIBUCIÓN BINOMIAL NEGATIVA O DE PASCAL.

$$f(x) = P(X = x) = \binom{x-1}{r-1} p^r q^{x-r} \quad ; \quad x = r, r+1, \ldots$$

$$\mu = \frac{r}{P} \quad ; \quad \sigma^2 = \frac{rq}{p^2} \quad ; \quad M(t) = \left(\frac{pe^t}{1 - qe^t}\right)^r$$

La variable aleatoria X representa el número de pruebas de Bernoulli hasta que ocurra el r-ésimo éxito, incluyendo la prueba en que ocurre éste. El caso especial $r = 1$ da la distribución geométrica.

3. LA DISTRIBUCIÓN EXPONENCIAL.

$$f(x) = \begin{cases} \alpha e^{-\alpha x} & x > 0 \\ 0 & x \leqq 0 \end{cases}$$

$$\mu = \frac{1}{\alpha} \quad ; \quad \sigma^2 = \frac{1}{\alpha^2} \quad ; \quad M(t) = \frac{\alpha}{\alpha - t}$$

4. LA DISTRIBUCIÓN DE WEIBULL.

$$f(x) = \begin{cases} abx^{b-1}e^{-ax^b} & x > 0 \\ 0 & x \leqq 0 \end{cases}$$

$$\mu = a^{-1/b}\Gamma\left(1 + \frac{1}{b}\right) \qquad \sigma^2 = a^{-2/b}\left[\Gamma\left(1 + \frac{2}{b}\right) - \Gamma^2\left(1 + \frac{1}{b}\right)\right]$$

5. LA DISTRIBUCIÓN DE MAXWELL.

$$f(x) = \begin{cases} \sqrt{2/\pi}\alpha^{3/2}x^2e^{-\alpha x^2/2} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

$$\mu = 2\sqrt{\frac{2}{\pi\alpha}} \quad ; \quad \sigma^2 = \left(3 - \frac{8}{\pi}\right)\alpha^{-1}$$

Problemas resueltos

LA DISTRIBUCIÓN BINOMIAL

4.1. Encuentre la probabilidad de que al lanzar tres veces una moneda balanceada aparezcan (*a*) 3 caras, (*b*) 2 sellos y 1 cara, (*c*) al menos 1 cara, (*d*) no más de un sello.

Método 1.

Sea H la letra que denote cara y T la letra que denote sello. Supongamos que por ejemplo designamos HTH para indicar cara en el primer lanzamiento, sello en el segundo y luego, cara en el tercero.

Dado que en cada lanzamiento pueden ocurrir dos posibilidades (cara o sello), hay un total de (2)(2)(2) = 8 resultados posibles, es decir, 8 puntos muestrales en el espacio muestral. Éstos son

$$HHH, HHT, HTH, HTT, TTH, THH, THT, TTT$$

Dado que tenemos una moneda balanceada, asignamos probabilidades iguales de 1/8 a cada uno. Por consiguiente,

(*a*) $P(\text{3 caras}) = P(HHH) = \frac{1}{8}$

(*b*) $P(\text{2 sellos y 1 cara}) = P(HTT \cup TTH \cup THT)$

$$= P(HTT) + P(TTH) + P(THT) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$$

(*c*) P(al menos una cara)

$= P(1, 2 \text{ ó } 3 \text{ caras})$

$= P(1 \text{ cara}) + P(2 \text{ caras}) + P(3 \text{ caras})$

$= P(HTT \cup THT \cup TTH) + P(HHT \cup HTH \cup THH) + P(HHH)$

$= P(HTT) + P(THT) + P(TTH) + P(HHT) + P(HTH) + P(THH) + P(HHH) = \frac{7}{8}$

De manea alterna,

$$P(\text{al menos 1 cara}) = 1 - P(\text{no caras}) = 1 - P(TTT) = 1 - \frac{1}{8} = \frac{7}{8}$$

(*d*) P(no más de un sello) $= P(0 \text{ sellos o } 1 \text{ sello})$

$= P(0 \text{ sellos}) + P(1 \text{ sello})$

$= P(HHH) + P(HHT \cup HTH \cup THH)$

$= P(HHH) + P(HHT) + P(HTH) + P(THH)$

$= \frac{4}{8} = \frac{1}{2}$

Método 2 (con fórmula).

(*a*) $P(3 \text{ caras}) = \binom{3}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^0 = \frac{1}{8}$

(*b*) $P(2 \text{ sellos y } 1 \text{ cara }) = \binom{3}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^1 = \frac{3}{8}$

(*c*) $P(\text{al menos una cara}) = P(1, 2 \text{ ó } 3 \text{ caras})$

$= P(1 \text{ cara}) + P(2 \text{ caras}) + P(3 \text{ caras})$

$= \binom{3}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^2 + \binom{3}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^1 + \binom{3}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^0 = \frac{7}{8}$

De manera alterna,

$P(\text{al menos 1 cara}) = 1 - P(\text{no caras})$

$= 1 - \binom{3}{0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^3 = \frac{7}{8}$

(*d*) P(no más de un sello) $= P(0 \text{ sellos o } 1 \text{ sello})$

$= P(0 \text{ sellos}) + P(1 \text{ sello})$

$= \binom{3}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^0 + \binom{3}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right) = \frac{1}{2}$

Debemos mencionar que también se puede usar la notación para variables aleatorias. Por ejemplo, si X es la variable aleatoria que denota el número de caras en 3 lanzamientos, (*c*) puede escribirse

$$P(\text{al menos 1 cara}) = P(X \geqq 1) = P(X = 1) + P(X = 2) + P(X = 3) = \frac{7}{8}$$

Usaremos ambos enfoques indistintamente.

4.2. Encuentre la probabilidad de que en cinco lanzamientos de un dado balanceado aparezca 3 (*a*) dos veces, (*b*) como máximo una vez, (*c*) al menos dos veces.

Sea X la variable aleatoria que indica el número de veces que aparece 3 en cinco lanzamientos de un dado balanceado. Tenemos

$$\text{Probabilidad de 3 en un lanzamiento sencillo} = p = \frac{1}{6}$$

$$\text{Probabilidad de no 3 en un lanzamiento sencillo} = q = 1 - p = \frac{5}{6}$$

(*a*) $P(\text{3 ocurra dos veces}) = P(X = 2) = \binom{5}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^3 = \frac{625}{3888}$

(*b*) $P(\text{3 ocurra máximo una vez}) = P(X \leqq 1) = P(X = 0) + P(X = 1)$

$$= \binom{5}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^5 + \binom{5}{1}\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^4$$

$$= \frac{3125}{7776} + \frac{3125}{7776} = \frac{3125}{3888}$$

(*c*) $P(\text{3 ocurra al menos dos veces})$

$$= P(X \geqq 2)$$

$$= P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)$$

$$= \binom{5}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^3 + \binom{5}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2 + \binom{5}{4}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^1 + \binom{5}{5}\left(\frac{1}{6}\right)^5\left(\frac{5}{6}\right)^0$$

$$= \frac{625}{3888} + \frac{125}{3888} + \frac{25}{7776} + \frac{1}{7776} = \frac{763}{3888}$$

4.3. Encuentre la probabilidad de que en una familia de 4 hijos haya (*a*) al menos un niño, (*b*) al menos 1 niño y al menos 1 niña. Suponga que la probabilidad de nacimiento de un varón es $1/2$.

(*a*) $P(\text{1 niño}) = \binom{4}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^3 = \frac{1}{4}$, $P(\text{2 niños}) = \binom{4}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^2 = \frac{3}{8}$

$P(\text{3 niños}) = \binom{4}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^1 = \frac{1}{4}$, $P(\text{4 niños}) = \binom{4}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^0 = \frac{1}{16}$

Entonces

$$P(\text{al menos un niño}) = P(\text{1 niño}) + P(\text{2 niños}) + P(\text{3 niños}) + P(\text{4 niños})$$

$$= \frac{1}{4} + \frac{3}{8} + \frac{1}{4} + \frac{1}{16} = \frac{15}{16}$$

Otro método

$$P(\text{al menos 1 niño}) = 1 - P(\text{no niño}) = 1 - \left(\frac{1}{2}\right)^4 = 1 - \frac{1}{16} = \frac{15}{16}$$

(*b*) $P(\text{al menos 1 niño y al menos 1 niña}) = 1 - P(\text{no niño}) - P(\text{no niña})$

$$= 1 - \frac{1}{16} - \frac{1}{16} = \frac{7}{8}$$

También hubiéramos podido resolver este problema colocando X como la variable aleatoria que denota el número de niños en familias con cuatro hijos. Entonces, por ejemplo, (*a*) se obtiene

$$P(X \geqq 1) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = \frac{15}{16}$$

4.4. De 2000 familias con 4 hijos cada una, cuántas esperaría encontrar con (*a*) al menos 1 niño, (*b*) 2 niños, (*c*) 1 ó 2 niñas, (*d*) no niñas.

Refiriéndonos al problema 4.3, vemos que

(*a*) Número esperado de familias con al menos 1 niño $= 2000\left(\frac{15}{16}\right) = 1875$

(*b*) Número esperado de familias con 2 niños $= 2000 \cdot P(\text{2 niños}) = 2000\left(\frac{3}{8}\right) = 750$

(*c*) $P(\text{1 ó 2 niñas}) = P(\text{1 niña}) + P(\text{2 niñas})$

$$= P(\text{1 niño}) + P(\text{2 niños}) = \frac{1}{4} + \frac{3}{8} = \frac{5}{8}$$

Número esperado de familias con 1 ó 2 niñas $= (2000)\left(\frac{5}{8}\right) = 1250$

(*d*) Número esperado de familias con no niñas $= (2000)\left(\frac{1}{16}\right) = 125$

4.5. Si el 20% de los tornillos producidos por una máquina son defectuosos, determine la probabilidad de que de 4 tornillos escogidos al azar, (*a*) 1, (*b*) 0, (*c*) menos de 2 tornillos sean defectuosos.

La probabilidad de un tornillo defectuoso es $p = 0.2$, la de un tornillo no defectuoso es $q = 1 - p = 0.8$. Sea la variable aleatoria X el número de tornillos defectuosos. Entonces

(*a*) $P(X = 1) = \binom{4}{1}(0.2)^1(0.8)^3 = 0.4096$

(*b*) $P(X = 0) = \binom{4}{0}(0.2)^0(0.8)^4 = 0.4096$

(*c*) $P(X < 2) = P(X = 0) + P(X = 1)$

$$= 0.4096 + 0.4096 = 0.8192$$

4.6. Encuentre la probabilidad de obtener, al menos una vez, un total de 7 en tres lanzamientos de un par de dados balanceados.

En un lanzamiento sencillo de un par de dados balanceados la probabilidad de 7 es $p = 1/6$ (*véase* problema 2.1, página 50), de manera que la probabilidad de no 7 en un lanzamiento sencillo es $q = 1 - p = 5/6$. Entonces

$$P(\text{no 7 en tres lanzamientos}) = \binom{3}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^3 = \frac{125}{216}$$

y

$$P(\text{al menos un 7 en tres lanzamientos}) = 1 - \frac{125}{216} = \frac{91}{216}$$

4.7. Encuentre la función generadora de momento de una variable aleatoria X que tiene distribución binomial.

Método 1.

Si X tiene distribución binomial,

$$f(x) = P(X = x) = \binom{n}{x}p^x q^{n-x}$$

Entonces la función generadora de momento está dada por

$$M(t) = E(e^{tx}) = \sum e^{tx} f(x)$$
$$= \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x q^{n-x}$$
$$= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x q^{n-x}$$
$$= (q + pe^t)^n$$

Método 2.

Para una secuencia de n pruebas de Bernoulli, defina

$$X_j = \begin{cases} 0 & \text{si hay error en la } j\text{-ésima prueba} \\ 1 & \text{si hay éxito en la } j\text{-ésima prueba} \end{cases} \qquad (j = 1, 2, \ldots, n)$$

Entonces las X_j son independientes y $X = X_1 + X_2 + \cdots + X_n$. Para la función generadora de momento de X_j tenemos

$$M_j(t) = e^{t0}q + e^{t1}p = q + pe^t \qquad (j = 1, 2, \ldots, n)$$

Entonces, según el teorema 3-9, página 87,

$$M(t) = M_1(t)M_2(t)\cdots M_n(t) = (q + pe^t)^n$$

4.8. Demuestre que la media y la varianza de una variable aleatoria con distribución binomial son, respectivamente, $\mu = np$ y $\sigma^2 = npq$.

Procediendo como en el método 2 del problema 4.7, tenemos para $j = 1, 2, \ldots, n$,

$$E(X_j) = 0q + 1p = p$$
$$\text{Var}(X_j) = E[(X_j - p)^2] = (0 - p)^2 q + (1 - p)^2 p$$
$$= p^2 q + q^2 p = pq(p + q) = pq$$

Entonces

$$\mu = E(X) = E(X_1) + E(X_2) + \cdots + E(X_n) = np$$
$$\sigma^2 = \text{Var}(X) = \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n) = npq$$

donde usamos el teorema 3-7 para σ^2.

También podemos obtener los resultados anteriores (pero con mayor dificultad) derivando la función generadora de momento (*véase* problema 3.38) o directamente a partir de la función de probabilidad.

4.9. Si la probabilidad de ocurrencia de un tornillo defectuoso es 0.1, encuentre (*a*) la media y (*b*) la desviación estándar, para el número de tornillos defectuosos de un total de 400 tornillos.

(*a*) Media $= \mu = np = (400)(0.1) = 40$, es decir, podemos esperar que haya 40 tornillos defectuosos

(*b*) Varianza $= \sigma^2 = npq = (400)(0.1)(0.9) = 36$. O sea, la desviación estándar $\sigma = \sqrt{36} = 6$.

LA LEY DE LOS GRANDES NÚMEROS PARA PRUEBAS DE BERNOULLI

4.10. Demuestre el teorema 4-1, la ley (débil) de los grandes números para pruebas de Bernoulli.

Según la desigualdad de Chebyshev, página 90, si X es una variable aleatoria con media finita μ y varianza σ^2, entonces,

(*1*) $$P(|X - \mu| \geqq k\sigma) \leqq \frac{1}{k^2}$$

En particular, si X tiene distribución binomial o de Bernoulli, entonces $\mu = np$, $\sigma = \sqrt{npq}$ y (*1*) se obtiene

$$P(|X - np| \geqq k\sqrt{npq}) \leqq \frac{1}{k^2} \tag{2}$$

o

$$P\left(\left|\frac{X}{n} - p\right| \geqq k\sqrt{\frac{pq}{n}}\right) \leqq \frac{1}{k^2} \tag{3}$$

Si remplazamos $\epsilon = k\sqrt{\dfrac{pq}{n}}$, (*3*) obtenemos

$$P\left(\left|\frac{X}{n} - p\right| \geqq \epsilon\right) \leqq \frac{pq}{n\epsilon^2}$$

y tomando el límite a medida que $n \to \infty$ tenemos,

$$\lim_{n\to\infty} P\left(\left|\frac{X}{n} - p\right| \geqq \epsilon\right) = 0$$

El resultado también se deduce directamente a partir del teorema 3-19, página 91, con $S_n = X$, $\mu = np$, $\sigma = \sqrt{npq}$.

4.11. Dé una interpretación de la ley (débil) de los grandes números para las ocurrencias de un 3 en lanzamientos sucesivos de un dado balanceado.

La ley de los grandes números plantea en este caso que la probabilidad de la proporción de 3 difiere 1/6 en n lanzamientos por más que cualquier valor $\epsilon > 0$ tiende a cero a medida que $n \to \infty$.

LA DISTRIBUCIÓN NORMAL

4.12. Encuentre el área bajo la curva normal estándar que aparece en la figura 4-3 (*a*) entre $z = 0$ y $z = 1.2$, (*b*) entre $z = -0.68$ y $z = 0$, (*c*) entre $z = -0.46$ y $z = 2.21$, (*d*) entre $z = 0.81$ y $z = 1.94$, (*e*) a la derecha de $z = -1.28$.

(*a*) Usando la tabla del apéndice C, desplácese hacia abajo por la columna denominada z hasta encontrar 1.2. Luego, desplácese hacia la derecha hasta la columna designada como 0. El resultado, 0.3849, es el área requerida y representa la probabilidad de que Z esté entre 0 y 1.2 (figura 4-3). Por consiguiente,

$$P(0 \leqq Z \leqq 1.2) = \frac{1}{\sqrt{2\pi}} \int_0^{1.2} e^{-u^2/2}\, du = 0.3849$$

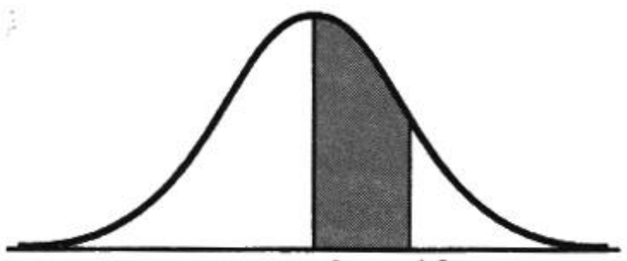


Fig. 4-3

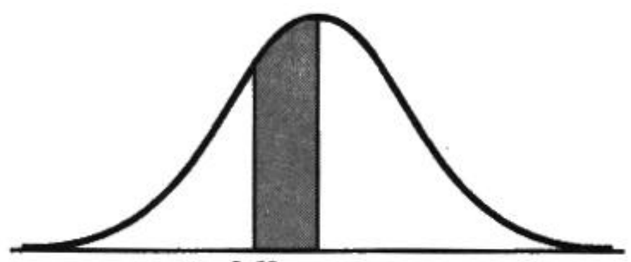


Fig. 4-4

(*b*) Área requerida = área entre $z = 0$ y $z = +0.68$ (por simetría). Por consiguiente, desplácese bajo la columna marcada como z hasta encontrar la entrada 0.6. Luego muévase hacia la derecha, hasta la columna marcada como 8.

El resultado, 0.2517, es el área requerida y representa la probabilidad de que Z esté entre -0.68 y 0 (figura 4-4). Por consiguiente,

$$P(-0.68 \leqq Z \leqq 0) = \frac{1}{\sqrt{2\pi}} \int_{-0.68}^{0} e^{-u^2/2}\, du$$

$$= \frac{1}{\sqrt{2\pi}} \int_0^{0.68} e^{-u^2/2}\, du = 0.2517$$

(*c*) Área requerida = (área entre $z = -0.46$ y $z = 0$)
+ (área entre $z = 0$ y $z = 2.21$)
= (área entre $z = 0$ y $z = 0.46$)
+ (área entre $z = 0$ y $z = 2.21$)
$= 0.1772 + 0.4864 = 0.6636$

El área, 0.6636, representa la probabilidad de que Z esté entre -0.46 y 2.21 (figura 4-5). Por consiguiente,

$$P(-0.46 \leqq Z \leqq 2.21) = \frac{1}{\sqrt{2\pi}} \int_{-0.46}^{2.21} e^{-u^2/2}\, du$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-0.46}^{0} e^{-u^2/2}\, du + \frac{1}{\sqrt{2\pi}} \int_{0}^{2.21} e^{-u^2/2}\, du$$

$$= \frac{1}{\sqrt{2\pi}} \int_{0}^{0.46} e^{-u^2/2}\, du + \frac{1}{\sqrt{2\pi}} \int_{0}^{2.21} e^{-u^2/2}\, du = 0.1772 + 0.4864$$

$$= 0.6636$$

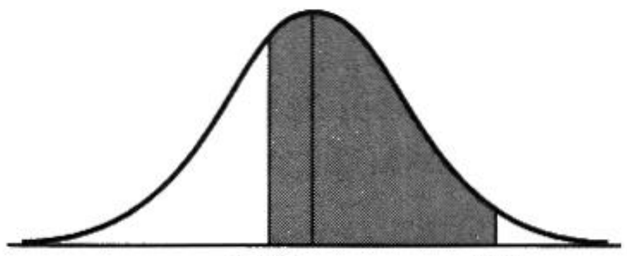


Fig. 4-5

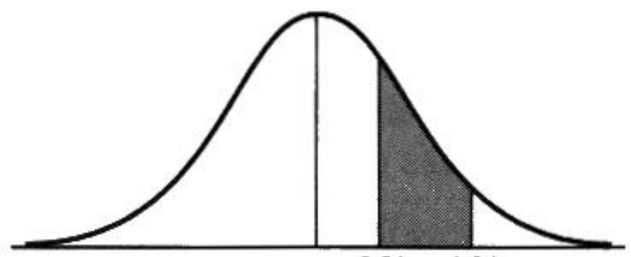


Fig. 4-6

(*d*) Área requerida (figura 4-6) = (área entre $z = 0$ y $z = 1.94$)
− (área entre $z = 0$ y $z = 0.81$)
$= 0.4738 - 0.2910 = 0.1828$

Esto es lo mismo que $P(0.81 \leqq Z \leqq 1.94)$.

(*e*) Área requerida (figura 4-7) = (área entre $z = -1.28$ y $z = 0$)
+ (área a la derecha de $z = 0$)
$= 0.3997 + 0.5 = 0.8997$

Esto es lo mismo que $P(Z \geqq -1.28)$

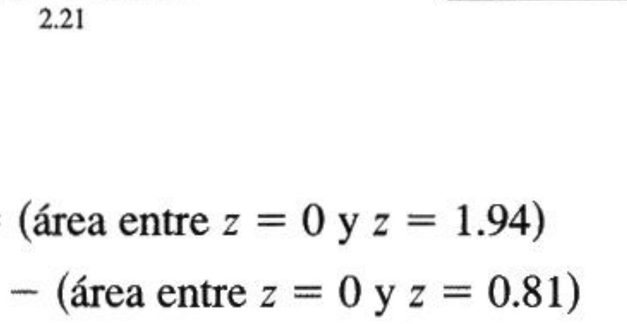


Fig. 4-7

4.13. Si "área" se refiere a lo que se encuentra bajo la curva normal estándar, encuentre el valor o valores de z tales que (*a*) área entre 0 y z sea 0.3770, (*b*) área a la izquierda de z sea 0.8621, (*c*) área entre -1.5 y z es 0.0217.

(*a*) En la tabla del apéndice C, la entrada 0.3770 está ubicada a la derecha de la fila marcada 1.1 y bajo la columna denominada 6. Entonces $z = 1.16$.

Por simetría, $z = -1.16$, es otro valor de z. Por consiguiente, $z = \pm 1.16$ (figura 4-8). El problema equivale a resolver la ecuación para z

$$\frac{1}{\sqrt{2\pi}} \int_0^z e^{-u^2/2}\, du = 0.3770$$

(*b*) Dado que el área es mayor que 0.5, z debe ser un valor positivo.

El área entre 0 y z es $0.8621 - 0.5 = 0.3621$, a partir de lo cual $z = 1.09$ (figura 4-9).

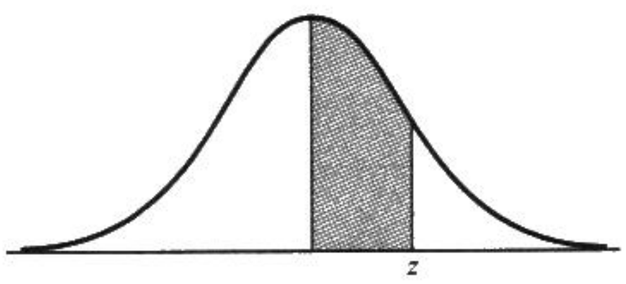


Fig. 4-8

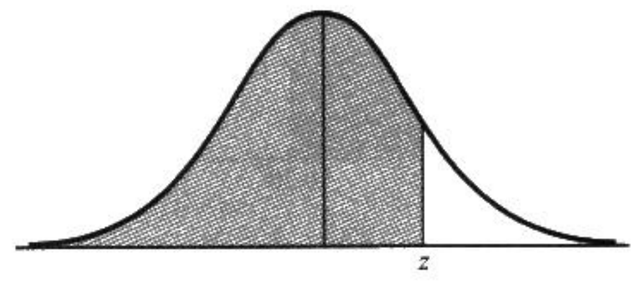


Fig. 4-9

(*c*) Si z es un valor positivo, el área será mayor que el área entre -1.5 y 0, la cual es 0.4332; así que z debe ser negativo.

Caso 1. z es negativo, pero a la derecha de -1.5 (figura 4-10).

$$\text{Área entre } -1.5 \text{ y } z = (\text{área entre } -1.5 \text{ y } 0) - (\text{área entre } 0 \text{ y } z)$$

$$0.0217 = 0.4332 - (\text{área entre } 0 \text{ y } z)$$

Entonces el área entre 0 y z es $0.4332 - 0.0217 - 0.4115$ a partir de lo cual $z = -1.35$.

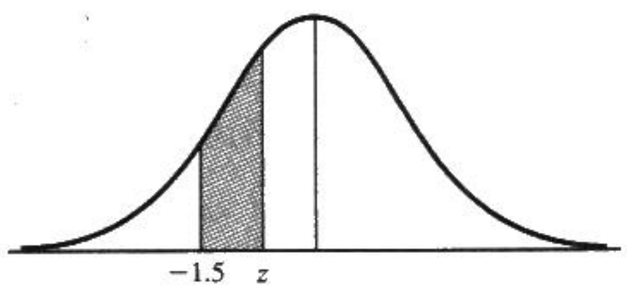


Fig. 4-10

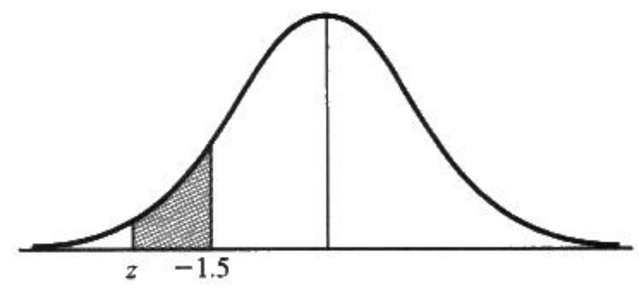


Fig. 4-11

Caso 2. z es negativo, pero a la izquierda de -1.5 (figura 4-11).

$$\text{Área entre } z \text{ y } -1.5 = (\text{área entre } z \text{ y } 0) - (\text{área entre } -1.5 \text{ y } 0)$$

$$0.0217 = (\text{área entre } 0 \text{ y } z) - 0.4332$$

Entonces, usando la interpolación lineal, el área entre 0 y z es $0.0217 + 0.4332 = 0.4549$ y $z = -1.694$; o con un poco menos de precisión, $z = -1.69$.

4.14. El peso promedio de 500 estudiantes varones en cierta universidad es de 151 lb y la desviación estándar es de 15 lb. Suponiendo que los pesos tienen distribución normal, encuentre cuántos estudiantes pesan (*a*) entre 120 y 155 lb, (*b*) más de 185 lb.

(*a*) Los pesos registrados entre 120 y 155 lb tienen en realidad cualquier valor entre 119.5 y 155.5 lb, suponiendo que se registraron con el valor más cercano a la libra (figura 4-12).

$$119.5 \text{ lb en unidades estándar} = (119.5 - 151)/15 = -2.10$$

$$155.5 \text{ lb en unidades estándar} = (155.5 - 151)/15 = 0.30$$

La proporción requerida de estudiantes = (área entre $z = -2.10$ y $z = 0.30$)
= (área entre $z = -2.10$ y $z = 0$)
+ (área entre $z = 0$ y $z = 0.30$)
= 0.4821 + 0.1179 = 0.6000

Entonces el número de estudiantes que pesan entre 120 y 155 lb es de 500(0.6000) = 300

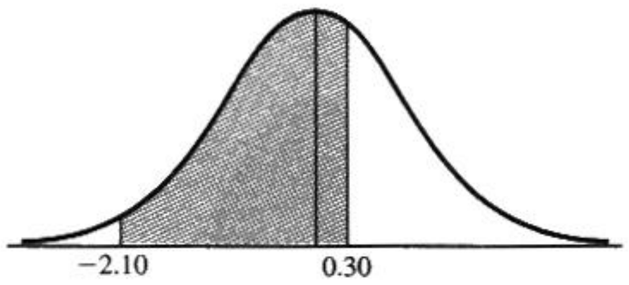


Fig. 4-12

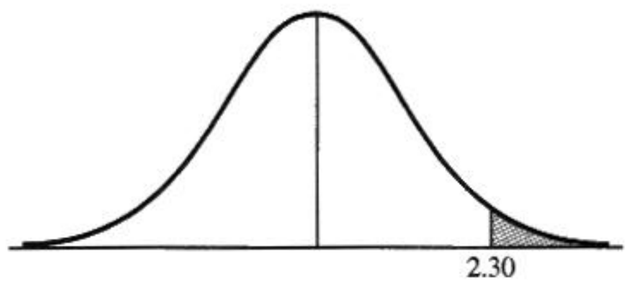


Fig. 4-13

(*b*) Los estudiantes que pesan más de 185 lb deben pesar al menos 185.5 lb (figura 4-13).

185.5 lb en unidades estándar = $(185.5 - 151)/15 = 2.30$

Proporción requerida de estudiantes
= (área a la derecha de $z = 2.30$)
= (área a la derecha de $z = 0$)
− (área entre $z = 0$ y $z = 2.30$)
= 0.5 − 0.4893 = 0.0107

Entonces el número de estudiantes que pesan más de 185 lb es 500(0.0107) = 5

Si W denota el peso de un estudiante escogido al azar, podemos resumir los resultados anteriores en términos de probabilidad escribiendo

$$P(119.5 \leqq W \leqq 155.5) = 0.6000 \qquad P(W \geqq 185.5) = 0.0107$$

4.15. La media del diámetro interior de una muestra de 200 empaques es 0.502 centímetros y la desviación estándar es 0.005 centímetros. El propósito para el cual se hicieron estos empaques permite un máximo de tolerancia en el diámetro de 0.496 a 0.508 centímetros, o de lo contrario se considera que los empaques son defectuosos. Determine el porcentaje de empaques defectuosos producidos por la máquina, suponiendo que los diámetros tienen distribución normal.

0.496 en unidades estándar = $(0.496 - 0.502)/0.005 = -1.2$

0.508 en unidades estándar = $(0.508 - 0.502)/0.005 = 1.2$

Proporción de empaques no defectuosos
= (área bajo la curva normal entre $z = -1.2$ y $z = 1.2$)
= (doble del área entre $z = 0$ y $z = 1.2$)
= 2(0.3849) = 0.7698, o 77%

Por consiguiente, el porcentaje de empaques defectuosos es 100% − 77% = 23% (figura 4-14).

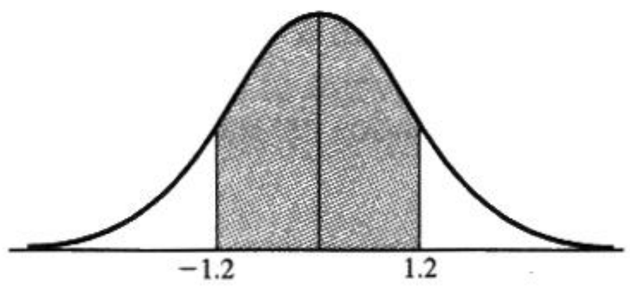


Fig. 4-14

Observe que si pensamos en los intervalos 0.496 a 0.508 centímetros como verdaderos representantes de los diámetros desde 0.4955 a 0.5085, el resultado anterior se modifica ligeramente. Sin embargo, para dos figuras significativas los resultados son los mismos.

4.16. Encuentre la función generadora de momento para la distribución normal general.

Tenemos

$$M(t) = E(e^{tX}) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-(x-\mu)^2/2\sigma^2}\, dx$$

Remplazando en la integral $(x - \mu)/\sigma = v$, de manera que $x = \mu + \sigma v$, $dx = \sigma dv$, tenemos

$$M(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\mu t + \sigma v t - (v^2/2)}\, dv = \frac{e^{\mu t + (\sigma^2 t^2/2)}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(v-\sigma t)^2/2}\, dv$$

Ahora remplazando $v - \sigma t = w$, encontramos que

$$M(t) = e^{\mu t + (\sigma^2 t^2/2)} \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-w^2/2}\, dw \right) = e^{\mu t + (\sigma^2 t^2/2)}$$

APROXIMACIÓN NORMAL A LA DISTRIBUCIÓN BINOMIAL

4.17. Encuentre la probabilidad de obtener entre 3 y 6 caras, inclusive, en 10 lanzamientos de una moneda balanceada usando (*a*) la distribución binomial, (*b*) la aproximación normal a la distribución binomial.

(*a*) Sea X la variable aleatoria que indica el número de caras que aparecen en 10 lanzamientos (figura 4-15). Entonces

$$P(X = 3) = \binom{10}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^7 = \frac{15}{128} \quad ; \quad P(X = 4) = \binom{10}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^6 = \frac{105}{512}$$

$$P(X = 5) = \binom{10}{5}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^5 = \frac{63}{256} \quad ; \quad P(X = 6) = \binom{10}{6}\left(\frac{1}{2}\right)^6\left(\frac{1}{2}\right)^4 = \frac{105}{512}$$

Entonces la probabilidad requerida es

$$P(3 \leqq X \leqq 6) = \frac{15}{128} + \frac{105}{512} + \frac{63}{256} + \frac{105}{512} = \frac{99}{128} = 0.7734$$

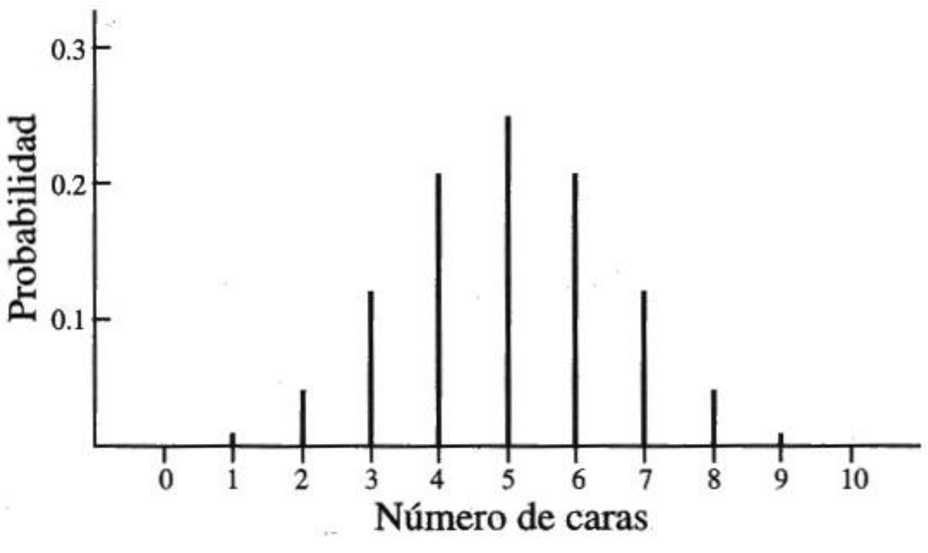


Fig. 4-15

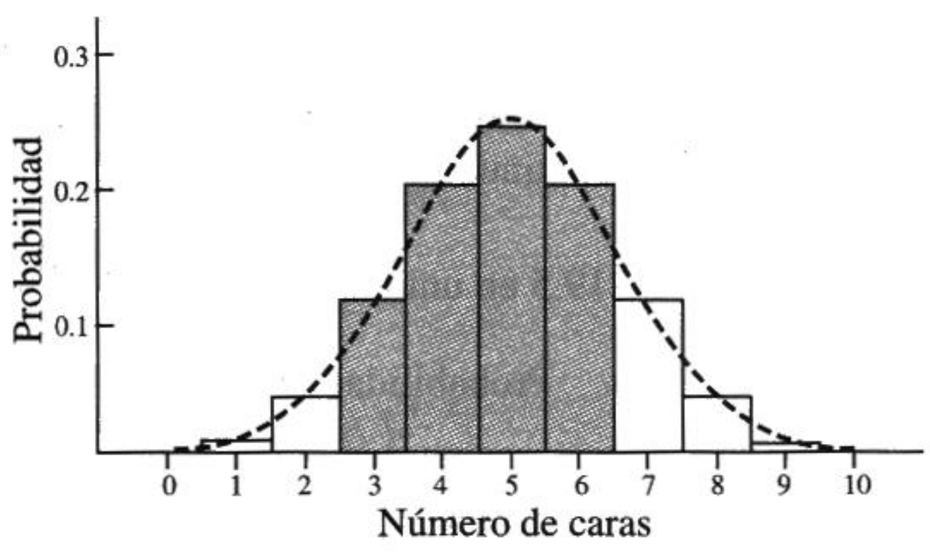


Fig. 4-16

(*b*) En las figuras 4-15 y 4-16 se muestra la distribución de probabilidad para el número de caras que aparecen en 10 lanzamientos de una moneda. La figura 4-16 trata los datos como si éstos fueran continuos. La probabilidad requerida es la suma de las áreas de los rectángulos sombreados de la figura 4-16 y puede aproximarse por el área bajo la correspondiente curva normal, que se muestra punteada. Al tratar los datos como continuos, se deduce que se pueden considerar de 3 a 6 caras como de 2.5 a 6.5 caras.

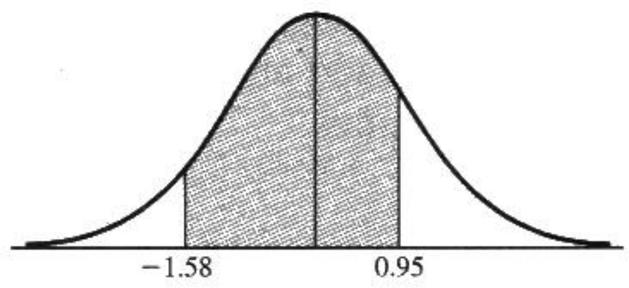


Fig. 4-17

También, la media y la varianza para la distribución binomial están dadas por $\mu = np = 10(\frac{1}{2}) = 5$ y $\sigma = \sqrt{npq} = \sqrt{(10)(\frac{1}{2})(\frac{1}{2})} = 1.58$.

ahora

$$2.5 \text{ en unidades estándar} = \frac{2.5 - 5}{1.58} = -1.58$$

$$6.5 \text{ en unidades estándar} = \frac{6.5 - 5}{1.58} = 0.95$$

Probabilidad requerida (figura 4-17)

$$= (\text{área entre } z = -1.58 \text{ y } z = 0.95)$$
$$= (\text{área entre } z = -1.58 \text{ y } z = 0)$$
$$+ (\text{área entre } z = 0 \text{ y } z = 0.95)$$
$$= 0.4429 + 0.3289 = 0.7718$$

lo cual se compara muy bien con el valor real 0.7734 obtenido en la parte (*a*). La precisión es aún mayor para valores grandes de *n*.

4.18. Se lanza una moneda balanceada 500 veces. Encuentre la probabilidad de que el número de caras no sea diferente de 250 por (*a*) más de 10, (*b*) más de 30.

$$\mu = np = (500)\left(\frac{1}{2}\right) = 250 \qquad \sigma = \sqrt{npq} = \sqrt{(500)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)} = 11.18$$

(*a*) Necesitamos que la probabilidad del número de caras esté entre 240 y 260, o si consideramos que los datos son continuos, entre 239.5 y 260.5.

$$239.5 \text{ en unidades estándar} = \frac{239.5 - 250}{11.18} = -0.94 \qquad 260.5 \text{ en unidades estándar} = 0.94$$

$$\text{Probabilidad requerida} = (\text{área bajo la curva normal entre } z = -0.94 \text{ y } z = 0.94)$$
$$= (\text{doble del área bajo la curva normal entre } z = 0 \text{ y } z = 0.94)$$
$$= 2(0.3264) = 0.6528$$

(*b*) Necesitamos que la probabilidad del número de caras esté entre 220 y 280, o si consideramos que los datos son continuos, entre 219.5 y 280.5.

$$219.5 \text{ en unidades estándar} = \frac{219.5 - 250}{11.18} = -2.73 \qquad 280.5 \text{ en unidades estándar} = 2.73$$

$$\text{Probabilidad requerida} = (\text{doble del área bajo la curva normal entre } z = 0 \text{ y } z = 2.73)$$
$$= 2(0.4968) = 0.9936$$

Se deduce que podemos tener mucha confianza en que el número de caras no difiera de lo esperado (250) por más de 30. Por consiguiente, si resulta que el número *real* de caras fue 280, podemos creer con gran certeza que la moneda no era balanceada, es decir, que estaba *cargada*.

4.19. Se lanza un dado 120 veces. Suponiendo que el dado está balanceado, encuentre la probabilidad de que 4 aparezca (*a*) 18 veces o menos, (*b*) 14 veces o menos.

La cara 4 tiene probabilidad $p = \frac{1}{6}$ de aparecer y probabilidad $q = \frac{5}{6}$ de no aparecer.

(*a*) Queremos obtener la probabilidad de que el número de cuatros caiga entre 0 y 18. Esto está dado exactamente por

$$\binom{120}{18}\left(\frac{1}{2}\right)^{18}\left(\frac{5}{6}\right)^{102}+\binom{120}{17}\left(\frac{1}{6}\right)^{17}\left(\frac{5}{6}\right)^{103}+\cdots+\binom{120}{0}\left(\frac{1}{6}\right)^{0}\left(\frac{5}{6}\right)^{120}$$

pero, dado que la labor de calcular es excesiva, usamos la aproximación normal.

Considerando los datos como continuos, se deduce que de 0 a 18 cuatros pueden tratarse como de -0.5 a 18.5 cuatros.

También

$$\mu = np = 120\left(\frac{1}{6}\right) = 20 \qquad \text{y} \qquad \sigma = \sqrt{npq} = \sqrt{(120)\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)} = 4.08$$

Entonces

$$-0.5 \text{ en unidades estándar} = \frac{-0.5-20}{4.08} = -5.02 \qquad 18.5 \text{ en unidades estándar} = -0.37$$

$$\begin{aligned}\text{Probabilidad requerida} &= (\text{área bajo la curva normal entre } z=-5.02 \text{ y } z=-0.37)\\ &= (\text{área entre } z=0 \text{ y } z=-5.02)\\ &\quad -(\text{área entre } z=0 \text{ y } z=-0.37)\\ &= 0.5-0.1443 = 0.3557\end{aligned}$$

(*b*) Procedemos como en la parte (*a*), remplazando 18 por 14. Entonces

$$-0.5 \text{ en unidades estándar} = -5.02 \qquad 14.5 \text{ en unidades estándar} = \frac{14.5-20}{4.08} = -1.35$$

$$\begin{aligned}\text{Probabilidad requerida} &= (\text{área bajo la curva normal entre } z=-5.02 \text{ y } z=-1.35)\\ &= (\text{área entre } z=0 \text{ y } z=-5.02)\\ &\quad -(\text{área entre } z=0 \text{ y } z=-1.35)\\ &= 0.5-0.4115 = 0.0885\end{aligned}$$

Se deduce que si fuéramos a tomar repetidamente muestras de 120 lanzamientos de un dado, un 4 aparecería 14 veces o menos en cerca de la décima parte de estas muestras.

LA DISTRIBUCIÓN DE POISSON

4.20. Establezca la validez de la aproximación de Poisson a la distribución binomial.

Si X tiene distribución binomial, entonces

$$P(X=x)=\binom{n}{x}p^x q^{n-x} \tag{1}$$

donde $E(X)=np$. Sea $\lambda = np$ de manera que $p=\lambda/n$. Entonces (1) se convierte en

$$\begin{aligned}P(X=x) &= \binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}\\ &= \frac{n(n-1)(n-2)\cdots(n-x+1)}{x!n^x}\lambda^x\left(1-\frac{\lambda}{n}\right)^{n-x}\\ &= \frac{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{x-1}{n}\right)}{x!}\lambda^x\left(1-\frac{\lambda}{n}\right)^{n-x}\end{aligned}$$

Ahora a medida que $n\to\infty$,

$$\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{x-1}{n}\right)\to 1$$

mientras

$$\left(1-\frac{\lambda}{n}\right)^{n-x}=\left(1-\frac{\lambda}{n}\right)^{n}\left(1-\frac{\lambda}{n}\right)^{-x}\to(e^{-\lambda})(1)=e^{-\lambda}$$

utilizando el bien conocido resultado del cálculo que dice

$$\lim_{n\to\infty}\left(1+\frac{u}{n}\right)^{n}=e^{u}$$

Se deduce que cuando $n\to\infty$ pero λ permanece fija (es decir, $p\to 0$),

$$P(X=x)\to\frac{\lambda^{x}e^{-\lambda}}{x!} \tag{2}$$

la cual es la distribución de Poisson.

Otro método.

La función generadora de momento para la distribución binomial es

$$(q+pe^{t})^{n}=(1-p+pe^{t})^{n}=[1+p(e^{t}-1)]^{n} \tag{3}$$

Si $\lambda=np$ de manera que $p=\lambda/n$, se obtiene

$$\left[1+\frac{\lambda(e^{t}-1)}{n}\right]^{n} \tag{4}$$

A medida que $n\to\infty$ esto se aproxima a

$$e^{\lambda(e^{t}-1)} \tag{5}$$

lo cual es la función generadora de momento de la distribución de Poisson. El resultado requerido se deduce usando el teorema 3-10 de la página 87.

4.21. Verifique que la función límite (2) del problema 4.20 es en realidad una función de probabilidad.

Primero, vemos que $P(X=x)>0$ para $x=0,1,\ldots$, dado que $\lambda>0$. Segundo, tenemos

$$\sum_{x=0}^{\infty}P(X=x)=\sum_{x=0}^{\infty}\frac{\lambda^{x}e^{-\lambda}}{x!}=e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^{x}}{x!}=e^{-\lambda}\cdot e^{\lambda}=1$$

y la verificación está completa.

4.22. El diez por ciento de las herramientas producidas en cierto proceso de fabricación son defectuosas. Encuentre la probabilidad de que en una muestra de 10 herramientas, seleccionadas al azar, exactamente 2 sean defectuosas usando (*a*) la distribución binomial, (*b*) la aproximación de Poisson a la distribución binomial.

(*a*) La probabilidad de que una herramienta sea defectuosa es $p=0.1$. Sea X el número de herramientas defectuosas de las 10 herramientas seleccionadas. Entonces, según la distribución binomial,

$$P(X=2)=\binom{10}{2}(0.1)^{2}(0.9)^{8}=0.1937\text{ ó }\ 0.19$$

(*b*) Tenemos $\lambda=np=(10)(0.1)=1$. Entonces, según la distribución de Poisson,

$$P(X=x)=\frac{\lambda^{x}e^{-\lambda}}{x!}\qquad\text{o}\qquad P(X=2)=\frac{(1)^{2}e^{-1}}{2!}=0.1839\text{ ó }\ 0.18$$

En general, la aproximación es buena si $p\leqq 0.1$ y $\lambda=np\leqq 5$.

4.23. Si la probabilidad de que una persona tenga una mala reacción a la inyección de determinado suero es 0.001, determine la probabilidad de que de cada 2000 individuos, (*a*) exactamente 3 y (*b*) más de 2 individuos tengan una reacción mala.

Sea X el número de individuos que reaccionan mal. X tiene distribución de Bernoulli, pero dado que suponemos que las reacciones malas son eventos raros, podemos suponer que X tiene distribución de Poisson, es decir,

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad \text{donde } \lambda = np = (2000)(0.001) = 2$$

(*a*)
$$P(X = 3) = \frac{2^3 e^{-2}}{3!} = 0.180$$

(*b*)
$$P(X > 2) = 1 - [P(X = 0) + P(X = 1) + P(X = 2)]$$
$$= 1 - \left[\frac{2^0 e^{-2}}{0!} + \frac{2^1 e^{-2}}{1!} + \frac{2^2 e^{-2}}{2!}\right]$$
$$= 1 - 5e^{-2} = 0.323$$

La evaluación exacta de las probabilidades usando la distribución binomial requerirá mucho más trabajo.

EL TEOREMA DEL LÍMITE CENTRAL

4.24. Verifique el teorema del límite central para una variable aleatoria X que tiene distribución binomial, y de ese modo establece la validez de la aproximación normal a la distribución binomial.

La variable estandarizada para X es $X^* = (X - np)/\sqrt{npq}$, y la función generadora de momento para X^* es

$$E(e^{tX^*}) = E(e^{t(X-np)/\sqrt{npq}})$$
$$= e^{-tnp/\sqrt{npq}} E(e^{tX/\sqrt{npq}})$$
$$= e^{-tnp/\sqrt{npq}} \sum_{x=0}^{n} e^{tx/\sqrt{npq}} \binom{n}{x} p^x q^{n-x}$$
$$= e^{-tnp/\sqrt{npq}} \sum_{x=0}^{n} \binom{n}{x} (pe^{t/\sqrt{npq}})^x q^{n-x}$$
$$= e^{-tnp/\sqrt{npq}} (q + pe^{t/\sqrt{npq}})^n$$
$$= [e^{-tp/\sqrt{npq}} (q + pe^{t/\sqrt{npq}})]^n$$
$$= (qe^{-tp/\sqrt{npq}} + pe^{tq/\sqrt{npq}})^n$$

Usando la expansión

$$e^u = 1 + u + \frac{u^2}{2!} + \frac{u^3}{3!} + \cdots$$

encontramos

$$qe^{-tp/\sqrt{npq}} + pe^{tq/\sqrt{npq}} = q\left(1 - \frac{tp}{\sqrt{npq}} + \frac{t^2p^2}{2npq} + \cdots\right)$$
$$+ p\left(1 + \frac{tq}{\sqrt{npq}} + \frac{t^2q^2}{2npq} + \cdots\right)$$
$$= q + p + \frac{pq(p+q)t^2}{2npq} + \cdots$$
$$= 1 + \frac{t^2}{2n} + \cdots$$

Por consiguiente,

$$E(e^{tX^*}) = \left(1 + \frac{t^2}{2n} + \cdots\right)^n$$

Pero a medida que $n \to \infty$ el lado derecho se aproxima a $e^{t^2/2}$, la cual es la función generadora de momento para la distribución normal estándar. Por consiguiente, el resultado requerido se deduce del teorema 3-10, página 87.

4.25. Demuestre el teorema del límite central (teorema 4-2, página 121).

Para $n = 1, 2, \ldots$, tenemos $S_n = X_1 + X_2 + \cdots + X_n$. Ahora, $X_1, X_2, \ldots, X_n$ tienen cada una media μ y varianza σ^2. Así,

$$E(S_n) = E(X_1) + E(X_2) + \cdots + E(X_n) = n\mu$$

y, dado que las X_k son independientes,

$$\text{Var}\,(S_n) = \text{Var}\,(X_1) + \text{Var}\,(X_2) + \cdots + \text{Var}\,(X_n) = n\sigma^2$$

Se deduce que la variable aleatoria estandarizada correspondiente a S_n es

$$S_n^* = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

La función generadora de momento para S_n^* es

$$\begin{aligned} E(e^{tS_n^*}) &= E[e^{t(S_n - n\mu)/\sigma\sqrt{n}}] \\ &= E[e^{t(X_1-\mu)/\sigma\sqrt{n}} e^{t(X_2-\mu)/\sigma\sqrt{n}} \cdots e^{t(X_n-\mu)/\sigma\sqrt{n}}] \\ &= E[e^{t(X_1-\mu)/\sigma\sqrt{n}}] \cdot E[e^{t(X_2-\mu)/\sigma\sqrt{n}}] \cdots E[e^{t(X_n-\mu)/\sigma\sqrt{n}}] \\ &= \{E[e^{t(X_1-\mu)/\sigma\sqrt{n}}]\}^n \end{aligned}$$

donde, en los dos últimos pasos, usamos respectivamente los hechos de que las X_k son independientes y tienen distribuciones idénticas. Ahora, según una expansión de series de Taylor,

$$\begin{aligned} E[e^{t(X_1-\mu)/\sigma\sqrt{n}}] &= E\left[1 + \frac{t(X_1-\mu)}{\sigma\sqrt{n}} + \frac{t^2(X_1-\mu)^2}{2\sigma^2 n} + \cdots\right] \\ &= E(1) + \frac{t}{\sigma\sqrt{n}} E(X_1 - \mu) + \frac{t^2}{2\sigma^2 n} E[(X_1-\mu)^2] + \cdots \\ &= 1 + \frac{t}{\sigma\sqrt{n}}(0) + \frac{t^2}{2\sigma^2 n}(\sigma^2) + \cdots = 1 + \frac{t^2}{2n} + \cdots \end{aligned}$$

de manera que

$$E(e^{tS_n^*}) = \left(1 + \frac{t^2}{2n} + \cdots\right)^n$$

Pero el límite de esto a medida que $n \to \infty$ es $e^{t^2/2}$, lo cual es la función generadora de momento de la distribución normal estandarizada. De ahí, según el teorema 3-10, página 87, se deduce el resultado requerido.

LA DISTRIBUCIÓN MULTINOMIAL

4.26. Una caja contiene 5 bolas rojas, 4 bolas blancas y 3 bolas azules. Se selecciona al azar una bola de la caja, se anota su color, y se devuelve a la caja. Encuentre la probabilidad de que de 6 bolas seleccionadas de esta manera, 3 sean rojas, 2 sean blancas y 1 sea azul.

Método 1 (por fórmula).

$$P(\text{rojo en cualquier retiro}) = \frac{5}{12} \qquad P(\text{blanco en cualquier retiro}) = \frac{4}{12}$$

$$P(\text{azul en cualquier retiro}) = \frac{3}{12}$$

Entonces

$$P(\text{ 3 rojas, 2 blancas, 1 azul}) = \frac{6!}{3!2!1!}\left(\frac{5}{12}\right)^3\left(\frac{4}{12}\right)^2\left(\frac{3}{12}\right)^1 = \frac{625}{5184}$$

Método 2.

La probabilidad de escoger cualquier bola roja es $5/12$. Entonces la probabilidad de escoger 3 bolas rojas es $(5/12)^3$. De manera similar, la probabilidad de escoger 2 bolas blancas es $(4/12)^2$, y de escoger 1 bola azul es $(3/12)^1$. Por consiguiente, la probabilidad de escoger 3 rojas, 2 blancas y 1 azul en tal orden es

$$\left(\frac{5}{12}\right)^3\left(\frac{4}{12}\right)^2\left(\frac{3}{12}\right)^1$$

Pero la misma selección puede lograrse en otras secuencias, y el número de ellas es

$$\frac{6!}{3!2!1!}$$

como se vio en el capítulo 1. Entonces, la probabilidad requerida es

$$\frac{6!}{3!2!1!}\left(\frac{5}{12}\right)^3\left(\frac{4}{12}\right)^2\left(\frac{3}{12}\right)^1$$

Método 3.

La probabilidad requerida es el término en $p_r^3 p_b^2 p_a$ en la expansión multinomial de $(P_r + P_a + P_b)^6$ donde $P_r = 5/12$, $P_b = 4/12$, $P_a = 3/12$. De la expansión real se obtiene el resultado anterior.

LA DISTRIBUCIÓN HIPERGEOMÉTRICA

4.27. Una caja contiene 6 canicas azules y 4 canicas rojas. Se lleva a cabo un experimento en el cual se escoge al azar una canica y se observa su color, pero la canica no se devuelve a la caja. Encuentre la probabilidad de que después de 5 pruebas del experimento, se hayan escogido 3 canicas azules.

Método 1.

El número de formas diferentes para seleccionar 3 canicas azules de 6 canicas azules es $\binom{6}{3}$. El número de formas diferentes para seleccionar las 2 canicas restantes de las 4 canicas rojas es $\binom{4}{2}$. Por consiguiente, el número de muestras diferentes que contienen 3 canicas azules y 2 canicas rojas es $\binom{6}{3}\binom{4}{2}$.

Ahora, el número total de formas diferentes de seleccionar 5 canicas de las 10 canicas $(6 + 4)$ que hay en la caja es $\binom{10}{5}$. Por lo tanto, la probabilidad requerida está dada por

$$\frac{\binom{6}{3}\binom{4}{2}}{\binom{10}{5}} = \frac{10}{21}$$

Método 2 (usando fórmula).

Tenemos $b = 6$, $r = 4$, $n = 5$, $x = 3$. Entonces según (*19*), página 122, la probabilidad requerida es

$$P(X = 3) = \frac{\binom{6}{3}\binom{4}{2}}{\binom{10}{2}}$$

LA DISTRIBUCIÓN UNIFORME

4.28. Demuestre que la media y la varianza de la distribución uniforme (página 123) están dadas respectivamente por (*a*) $\mu = (a + b)/2$, (*b*) $\sigma^2 = (b - a)^2/12$.

(*a*)
$$\mu = E(X) = \int_a^b \frac{x\,dx}{b-a} = \frac{x^2}{2(b-a)}\bigg|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}$$

(*b*) Tenemos

$$E(X^2) = \int_a^b \frac{x^2\,dx}{b-a} = \frac{x^3}{3(b-a)}\bigg|_a^b = \frac{b^3-a^3}{3(b-a)} = \frac{b^2+ab+a^2}{3}$$

La varianza está dada por

$$\sigma^2 = E[(X-\mu)^2] = E(X^2) - \mu^2$$
$$= \frac{b^2+ab+a^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{1}{12}(b-a)^2$$

LA DISTRIBUCIÓN DE CAUCHY

4.29. Demuestre que (*a*) la función generadora de momento para una variable aleatoria X con distribución de Cauchy no existe, pero que (*b*) la función característica sí existe.

(*a*) La función generadora de momento de X es

$$E(e^{tX}) = \frac{a}{\pi}\int_{-\infty}^{\infty} \frac{e^{tx}}{x^2+a^2}\,dx$$

la cual no existe si t es real. Esto puede verse, por ejemplo, al observar que si $x \geqq 0$, $t > 0$,

$$e^{tx} = 1 + tx + \frac{t^2x^2}{2!} + \cdots > \frac{t^2x^2}{2}$$

de manera que

$$\frac{a}{\pi}\int_{-\infty}^{\infty} \frac{e^{tx}}{x^2+a^2}\,dx \geqq \frac{at^2}{2\pi}\int_0^{\infty} \frac{x^2}{x^2+a^2}\,dx$$

y la integral a la derecha diverge.

(*b*) La función característica de X es

$$E(e^{i\omega X}) = \frac{a}{\pi}\int_{-\infty}^{\infty} \frac{e^{i\omega x}}{x^2+a^2}\,dx$$
$$= \frac{a}{\pi}\int_{-\infty}^{\infty} \frac{\cos\omega x}{x^2+a^2}\,dx + \frac{ai}{\pi}\int_{-\infty}^{\infty} \frac{\operatorname{sen}\omega x}{x^2+a^2}\,dx$$
$$= \frac{2a}{\pi}\int_0^{\infty} \frac{\cos\omega x}{x^2+a^2}\,dx$$

donde hemos usado el hecho de que los integrandos en el penúltimo renglón son funciones par e impar, respectivamente. Se puede demostrar que la última integral existe y es igual a e^{-aw}.

4.30. Sea Θ una variable aleatoria distribuida uniformemente en el intervalo $-\frac{\pi}{2} \leqq \theta \leqq \frac{\pi}{2}$. Demuestre que $X = a\tan\Theta$, $a > 0$, tiene distribución de Cauchy en $-\infty < x < \infty$.

La función de densidad de Θ es

$$f(\theta) = \frac{1}{\pi} \qquad -\frac{\pi}{2} \leqq \theta \leqq \frac{\pi}{2}$$

Considerando la transformación $x = a\tan\theta$, tenemos

$$\theta = \tan^{-1}\frac{x}{a} \qquad \text{y} \qquad \frac{d\theta}{dx} = \frac{a}{x^2+a^2} > 0$$

Entonces, según el teorema 2-3, página 47, la función de densidad de X está dada por

$$g(x) = f(\theta)\left|\frac{d\theta}{dx}\right| = \frac{1}{\pi}\frac{a}{x^2 + a^2}$$

la cual es la distribución de Cauchy.

LA DISTRIBUCIÓN GAMMA

4.31. Demuestre que la media y la varianza de la distribución gamma están dadas por (*a*) $\mu = \alpha\beta$, (*b*) $\sigma^2 = \alpha\beta^2$.

(*a*)
$$\mu = \int_0^\infty x\left[\frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)}\right]dx = \int_0^\infty \frac{x^\alpha e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)}\,dx$$

Remplazando $x/\beta = t$, tenemos

$$\mu = \frac{\beta^\alpha\beta}{\beta^\alpha\Gamma(\alpha)}\int_0^\infty t^\alpha e^{-t}\,dt = \frac{\beta}{\Gamma(\alpha)}\Gamma(\alpha+1) = \alpha\beta$$

(*b*)
$$E(X^2) = \int_0^\infty x^2\left[\frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)}\right]dx = \int_0^\infty \frac{x^{\alpha+1}e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)}\,dx$$

Remplazando $x/\beta = t$, tenemos

$$E(X^2) = \frac{\beta^{\alpha+1}\beta}{\beta^\alpha\Gamma(\alpha)}\int_0^\infty t^{\alpha+1}e^{-t}\,dt$$

$$= \frac{\beta^2}{\Gamma(\alpha)}\Gamma(\alpha+2) = \beta^2(\alpha+1)\alpha$$

dado que $\Gamma(\alpha+2) = (\alpha+1)\Gamma(\alpha+1) = (\alpha+1)\alpha\Gamma(\alpha)$. Por consiguiente,

$$\sigma^2 = E(X^2) - \mu^2 = \beta^2(\alpha+1)\alpha - (\alpha\beta)^2 = \alpha\beta^2$$

LA DISTRIBUCIÓN BETA

4.32. Encuentre la media de la distribución beta.

$$\mu = E(X) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 x[x^{\alpha-1}(1-x)^{\beta-1}]\,dx$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 x^\alpha(1-x)^{\beta-1}\,dx$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+1+\beta)}$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\alpha\Gamma(\alpha)\Gamma(\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta)} = \frac{\alpha}{\alpha+\beta}$$

4.33. Encuentre la varianza de la distribución beta.

El segundo momento alrededor del origen es

$$E(X^2) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^2[x^{\alpha-1}(1-x)^{\beta-1}]\,dx$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{\alpha+1}(1-x)^{\beta-1}\,dx$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+2)\Gamma(\beta)}{\Gamma(\alpha+2+\beta)}$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{(\alpha+1)\alpha\Gamma(\alpha)\Gamma(\beta)}{(\alpha+\beta+1)(\alpha+\beta)\Gamma(\alpha+\beta)}$$

$$= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}$$

Entonces, usando el problema 4.32, la varianza es

$$\sigma^2 = E(X^2) - [E(X)]^2 = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} - \left(\frac{\alpha}{\alpha+\beta}\right)^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

LA DISTRIBUCIÓN CHI CUADRADO

4.34. Demuestre que la función generadora de momento de una variable aleatoria X, la cual tiene distribución chi cuadrado con ν grados de libertad, es $M(t) = (1-2t)^{-\nu/2}$.

$$M(t) = E(e^{tX}) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \int_0^\infty e^{tx} x^{(\nu-2)/2} e^{-x/2}\,dx$$

$$= \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \int_0^\infty x^{(\nu-2)/2} e^{-(1-2t)x/2}\,dx$$

Remplazando en la última integral $(1-2t)x/2 = u$, encontramos

$$M(t) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \int_0^\infty \left(\frac{2u}{1-2t}\right)^{(\nu-2)/2} e^{-u} \frac{2\,du}{1-2t}$$

$$= \frac{(1-2t)^{-\nu/2}}{\Gamma(\nu/2)} \int_0^\infty u^{(\nu/2)-1} e^{-u}\,du = (1-2t)^{-\nu/2}$$

4.35. Sean X_1 y X_2 variables aleatorias independientes con distribución chi cuadrado con ν_1 y ν_2 grados de libertad, respectivamente. (*a*) Demuestre que la función generadora de momento de $Z = X_1 + X_2$ es $(1-2t)^{-(\nu_1+\nu_2)/2}$, de ahí, (*b*) demuestre que Z tiene distribución chi cuadrado con $\nu_1 + \nu_2$ grados de libertad.

(*a*) La función generadora de momento de $Z = X_1 + X_2$ es

$$M(t) = E[e^{t(X_1+X_2)}] = E(e^{tX_1})E(e^{tX_2}) = (1-2t)^{-\nu_1/2}(1-2t)^{-\nu_2/2} = (1-2t)^{-(\nu_1+\nu_2)/2}$$

usando el problema 4.34.

(*b*) Se puede ver a partir del problema 4.34 que una distribución cuya función generadora de momento es $(1-2t)^{-(\nu_1+\nu_2)/2}$ es la distribución chi cuadrado con $\nu_1 + \nu_2$ grados de libertad. Esto debe ser la distribución de Z, según el teorema 3-10, página 87.

Generalizando los resultados anteriores, obtenemos una demostración del teorema 4-4, página 125.

4.36. Sea X una variable aleatoria distribuida normalmente con media 0 y varianza 1. Demuestre que X^2 tiene distribución chi cuadrado con 1 grado de libertad.

Queremos encontrar la distribución de $Y = X^2$ dada una distribución normal para X. Dado que la correspondencia entre X y Y no es uno a uno, no podemos aplicar el teorema 2-3 tal como está, pero podemos proceder de la siguiente manera.

Para $y < 0$, es claro que $P(Y \leqq y) = 0$. Para $y \geqq 0$, tenemos

$$P(Y \leqq y) = P(X^2 \leqq y) = P(-\sqrt{y} \leqq X \leqq +\sqrt{y})$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{y}}^{+\sqrt{y}} e^{-x^2/2}\, dx = \frac{2}{\sqrt{2\pi}} \int_{0}^{+\sqrt{y}} e^{-x^2/2}\, dx$$

donde el último paso utiliza el hecho de que la función de densidad normal estándar es simétrico. Cambiando en la integral la variable $x = +\sqrt{t}$, obtenemos

$$P(Y \leqq y) = \frac{1}{\sqrt{2\pi}} \int_{0}^{y} t^{-1/2} e^{-t/2}\, dt$$

Pero ésta es una distribución chi cuadrado con 1 grado de libertad, como se observa al colocar $v = 1$ en (*39*), página 124, y al usar el hecho de que $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

4.37. Demuestre el teorema 4-3, página 121, para $v = 2$.

Del problema 4.36 vemos que si X_1 y X_2 tienen distribución normal con media 0 y varianza 1, entonces X_1^2 y X_2^2 tienen distribución chi cuadrado con 1 grado de libertad cada una. Entonces, a partir del problema 4.35(*b*), vemos que $Z = X_1^2 + X_2^2$ tiene distribución chi cuadrado con $1 + 1 = 2$ grados de libertad si X_1 y X_2 son independientes. El resultado general para todos los enteros positivos v se deduce de igual manera.

4.38. En la figura 4-18 se muestra la gráfica de la distribución chi cuadrado con 5 grados de libertad (*véase* los comentarios sobre la notación en la página 125). Encuentre los valores χ_1^2, χ_2^2 para los cuales

(*a*) el área sombreada a la derecha = 0.05

(*b*) el área total sombreada = 0.05

(*c*) el área sombreada a la izquierda = 0.10

(*d*) el área sombreada a la derecha = 0.01

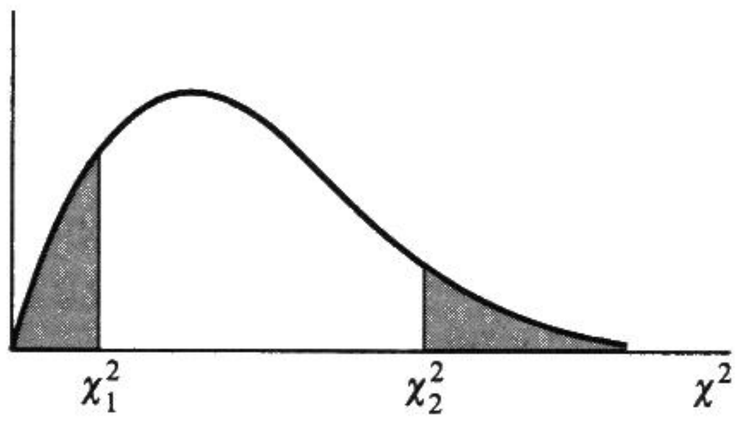


Fig. 4-18

(*a*) Si el área sombreada a la derecha es 0.05, entonces el área a la izquierda de χ_2^2 es $(1 - 0.05) = 0.95$, y χ_2^2 representa el percentil 95, $\chi_{0.95}^2$.

En relación a la tabla del apéndice E, desplácese por la columna v hasta encontrar la entrada 5. Luego, desplácese hacia la derecha, hacia la columna llamada $\chi_{0.95}^2$. El resultado, 11,1, es el valor requerido para χ^2.

(*b*) Como la distribución no es simétrica, hay muchos valores para los cuales el área sombreada total = 0.05. Por ejemplo, el área sombreada a la derecha puede ser 0.04, mientras que el área sombreada a la izquierda es 0.01. Sin embargo, se acostumbra escoger dos áreas de igual tamaño, a menos que se especifique lo contrario. Entonces, en este caso, cada área = 0.025.

Si el área sombreada a la derecha es 0.025, el área a la izquierda de χ_2^2 es 1 2 0.025 5 0.975 y χ_2^2 representa el percentil 97.5, $\chi_{0.975}^2$, el cual es 12.8, según el apéndice E.

De manera similar, si el área sombreada a la izquierda de 0.025, el área a la izquierda de χ_1^2 es 0.025 y χ_1^2 representa el percentil 2.5, $\chi_{0.025}^2$, el cual es igual a 0.831.

Por consiguiente, los valores son 0.831 y 12.8.

(*c*) Si el área sombreada a la izquierda es 0.10, χ_1^2 representa el décimo percentil, $\chi_{0.10}^2$, el cual es igual a 1,61.

(*d*) Si el área sombreada a la derecha es 0.01, el área a la izquierda de χ_2^2 es 0.99 y χ_2^2 representa el percentil 99, $\chi_{0.99}^2$, el cual es igual a 15.1.

4.39. Encuentre los valores de χ^2 para los que el área de la cola a mano derecha de la distribución χ^2 es 0.05, si el número de grados de libertad v es igual a (*a*) 15, (*b*) 21, (*c*) 50.

Usando la tabla del apéndice E, encontramos en la columna titulada $\chi_{0.95}^2$ los valores: (*a*) 25.0 correspondiente a $v = 15$; (*b*) 32.7 correspondiente a $v = 21$; (*c*) 67.5 correspondiente a $v = 50$.

4.40. Encuentre el valor de la mediana de χ^2 correspondiente a (*a*) 9, (*b*) 28, (*c*) 40 grados de libertad.

Usando la tabla del apéndice E, encontramos en la columna titulada $\chi_{0.50}^2$ (dado que la media es el percentil 50) los valores: (*a*) 8.34 correspondiente a $v = 9$; (*b*) 27.3 correspondiente a $v = 28$; (*c*) 39.3 correspondiente a $v = 40$.

Es interesante anotar que los valores para la mediana son casi iguales al número de grados de libertad. De hecho, para $v > 10$ los valores de la mediana son iguales a $v - 0.7$, como se puede ver en la tabla.

4.41. Encuentre $\chi_{0.95}^2$ para (*a*) $v = 50$, (*b*) $v = 100$ grados de libertad.

Para v mayor que 30 podemos usar el hecho de que $(\sqrt{2\chi^2} - \sqrt{2\nu - 1})$ está muy cercano a una distribución normal con media cero y varianza uno. Entonces, si z_p es el $(100p)$ percentil de la distribución normal estandarizada, podemos escribir, con un alto grado de aproximación,

$$\sqrt{2\chi_p^2} - \sqrt{2\nu - 1} = z_p \qquad \text{o} \qquad \sqrt{2\chi_p^2} = z_p + \sqrt{2\nu - 1}$$

a partir de lo cual

$$\chi_p^2 = \tfrac{1}{2}(z_p + \sqrt{2\nu - 1})^2$$

(*a*) Si $\nu = 50$, $\chi_{0.95}^2 = \frac{1}{2}(z_{0.95} + \sqrt{2(50) - 1})^2 = \frac{1}{2}(1.64 + \sqrt{99})^2 = 69.2$, el cual está muy de acuerdo con el valor 67.5 dado en el apéndice E.

(*b*) Si $\nu = 100$, $\chi_{0.95}^2 = \frac{1}{2}(z_{0.95} + \sqrt{2(100) - 1})^2 = \frac{1}{2}(1.64 + \sqrt{199})^2 = 124.0$ (valor real = 124.3).

LA DISTRIBUCIÓN *t* DE STUDENT

4.42. Demuestre el teorema 4-6, página 126.

Dado que Y tiene distribución normal con media 0 y varianza 1, su función de densidad es

$$(1) \qquad \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$$

Dado que Z tiene distribución chi cuadrado con v grados de libertad, su función de densidad es

$$(2) \qquad \frac{1}{2^{\nu/2}\Gamma(\nu/2)} z^{(\nu/2)-1} e^{-z/2} \qquad z > 0$$

Debido a que Y y Z son independientes, su función de densidad conjunta es el producto de (*1*) y (*2*), o sea,

$$\frac{1}{\sqrt{2\pi}2^{\nu/2}\Gamma(\nu/2)} z^{(\nu/2)-1} e^{-(y^2+z)/2}$$

para $-\infty < y < +\infty$, $z > 0$.

La función de distribución de $T = Y/\sqrt{Z/\nu}$ es

$$F(x) = P(T \leqq x) = P(Y \leqq x\sqrt{Z/\nu})$$
$$= \frac{1}{\sqrt{2\pi}2^{\nu/2}\Gamma(\nu/2)} \iint_{\mathscr{R}} z^{(\nu/2)-1} e^{-(y^2+z)/2}\, dy\, dz$$

donde la integral se toma sobre la región $\mathscr{R}$ del plano zy para le cual $y \leqq x\sqrt{z/\nu}$. Primero fijamos z e integramos con respecto a y desde $-\infty$ hasta $x\sqrt{z/\nu}$. Luego integramos con respecto a z desde 0 hasta ∞. Por lo tanto, tenemos

$$F(x) = \frac{1}{\sqrt{2\pi}\,2^{\nu/2}\Gamma(\nu/2)} \int_{z=0}^{\infty} z^{(\nu/2)-1} e^{-z/2} \left[\int_{y=-\infty}^{x\sqrt{z/\nu}} e^{-y^2/2}\, dy \right] dz$$

Remplazando $y = u\sqrt{z/\nu}$ en la integral que está entre paréntesis rectangulares, encontramos

$$F(x) = \frac{1}{\sqrt{2\pi}\,2^{\nu/2}\Gamma(\nu/2)} \int_{z=0}^{\infty} \int_{u=-\infty}^{\infty} z^{(\nu/2)-1} e^{-z/2} \sqrt{z/\nu}\, e^{-u^2 z/2\nu}\, du\, dz$$
$$= \frac{1}{\sqrt{2\pi\nu}\,2^{\nu/2}\Gamma(\nu/2)} \int_{u=-\infty}^{x} \left[\int_{z=0}^{\infty} z^{(\nu-1)/2} e^{-(z/2)[1+(u^2/\nu)]}\, dz \right] du$$

Remplazando $w = \frac{z}{2}\left(1 + \frac{u^2}{\nu}\right)$ esto puede escribirse

$$F(x) = \frac{1}{\sqrt{2\pi\nu}\,2^{\nu/2}\Gamma(\nu/2)} \cdot 2^{(\nu+1)/2} \int_{u=-\infty}^{x} \left[\int_{w=0}^{\infty} \frac{w^{(\nu-1)/2} e^{-w}}{(1 + u^2/\nu)^{(\nu+1)/2}}\, dw \right] du$$
$$= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)} \int_{u=-\infty}^{x} \frac{du}{(1 + u^2/\nu)^{(\nu+1)/2}}$$

como se requiere.

4.43. La gráfica de la distribución t de Student con 9 grados de libertad se muestra en la figura 4-19. Encuentre el valor t_1 para el cual

(*a*) el área sombreada a la derecha = 0.05,

(*b*) el área sombreada total = 0.05,

(*c*) el área total sin sombrear = 0.99,

(*d*) el área sombreada a la izquierda = 0.01,

(*e*) el área a la izquierda de t_1 es 0.90.

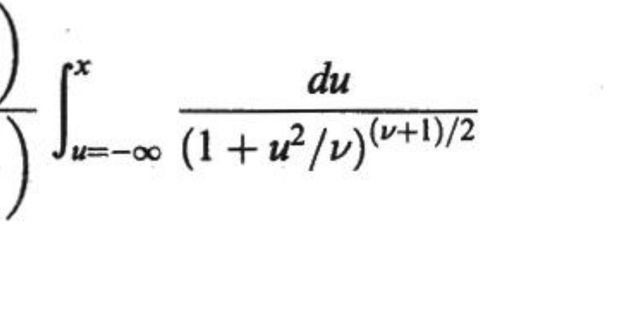


Fig. 4-19

(*a*) Si el área sombreada a la derecha es 0.05, entonces el área a la izquierda de t_1 es $(1 - 0.05) = 0.95$, y t_1 representa el percentil 95, $t_{0.95}$.

Refiriéndonos a la tabla del apéndice D, desplácese hacia abajo por la columna titulada v hasta llegar a la entrada 9. Luego diríjase hacia la derecha hasta la columna llamada $t_{0.95}$. El resultado 1.83 es el valor requerido de t.

(*b*) Si el área sombreada total es 0.05, entonces, por simetría, el área sombreada a la derecha es 0.025. Por consiguiente, el área a la izquierda de t_1 es $(1 - 0.025) = 0.975$, y t_1 representa el percentil 97.5, $t_{0.975}$. En el apéndice D encontramos que 2.26 es el valor requerido de t.

(*c*) Si el área total sin sombrear es 0.99, entonces el área total sombreada es $(1 - 0.99) = 0.01$, y el área sombreada a la derecha es $0.01/2 = 0.005$. En la tabla encontramos que $t_{0.995} = 3.25$.

(*d*) Si el área sombreada a la izquierda es 0.01, entonces, por simetría, el área sombreada a la derecha es 0.01. De la tabla, $t_{0.99} = 2.82$. Por consiguiente, el valor t para el cual el área sombreada de la izquierda es 0.01 es -2.82.

(*e*) Si el área a la izquierda de t_1 es 0.90, entonces t_1 corresponde al percentil 90, $t_{0.995}$ el cual por la tabla es igual a 1.38.

4.44. Encuentre los valores de t para los cuales el área de la cola a la derecha de la distribución t es 0.05 si el número de grados de libertad v es igual a (*a*) 16, (*b*) 27, (*c*) 200.

Refiriéndonos al apéndice D, encontramos en la columna llamada $t_{0.95}$ los valores (*a*) 1.75 correspondiente a $v = 16$; (*b*) 1.70 correspondiente a $v = 27$; (*c*) 1.645 correspondiente a $v = 200$ (el último es el valor que se puede obtener usando la curva normal. En el apéndice D, este valor corresponde a la entrada en la última fila marcada como ∞).

LA DISTRIBUCIÓN *F*

4.45. Demuestre el teorema 4-7.

La distribución conjunta de V_1 y V_2 está dada por

$$f(v_1, v_2) = \left(\frac{1}{2^{\nu_1/2}\Gamma(\nu_1/2)}\, v_1^{(\nu_1/2)-1} e^{-v_1/2}\right)\left(\frac{1}{2^{\nu_2/2}\Gamma(\nu_2/2)}\, v_2^{(\nu_2/2)-1} e^{-v_2/2}\right)$$

$$= \frac{1}{2^{(\nu_1+\nu_2)/2}\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\, v_1^{(\nu_1/2)-1} v_2^{(\nu_2/2)-1} e^{-(v_1+v_2)/2}$$

si $v_1 > 0$, $v_2 > 0$ y 0 de otra manera. Haga la transformación

$$u = \frac{v_1/\nu_1}{v_2/\nu_2} = \frac{\nu_2 v_1}{\nu_1 v_2}, \qquad w = v_2 \qquad \text{o} \qquad v_1 = \frac{\nu_1 uw}{\nu_2} \qquad v_2 = w$$

Entonces el jacobiano es

$$\frac{\partial(v_1, v_2)}{\partial(u, w)} = \begin{vmatrix} \partial v_1/\partial u & \partial v_1/\partial w \\ \partial v_2/\partial u & \partial v_2/\partial w \end{vmatrix} = \begin{vmatrix} \nu_1 w/\nu_2 & \nu_1 u/\nu_2 \\ 0 & 1 \end{vmatrix} = \frac{\nu_1 w}{\nu_2}$$

Denotando la densidad como una función de u y w por $g(u, w)$, entonces tenemos

$$g(u, w) = \frac{1}{2^{(\nu_1+\nu_2)/2}\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \left(\frac{\nu_1 uw}{\nu_2}\right)^{(\nu_1/2)-1} w^{(\nu_2/2)-1} e^{-[1+(\nu_1 u/\nu_2)](w/2)}\, \frac{\nu_1 w}{\nu_2}$$

si $u > 0$, $w > 0$ y 0 de otra manera.

La función de densidad (marginal) de U puede encontrarse integrando con respecto a w desde 0 hasta ∞, es decir,

$$h(u) = \frac{(\nu_1/\nu_2)^{\nu_1/2} u^{(\nu_1/2)-1}}{2^{(\nu_1+\nu_2)/2}\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \int_0^\infty w^{[(\nu_1+\nu_2)/2]-1} e^{-[1+(\nu_1 u/\nu_2)](w/2)}\, dw$$

si $u > 0$ y 0 si $u \leqq 0$. Pero a partir de 15, apéndice A,

$$\int_0^\infty w^{p-1} e^{-aw}\, dw = \frac{\Gamma(p)}{a^p}$$

Por consiguiente, tenemos

$$h(u) = \frac{(\nu_1/\nu_2)^{\nu_1/2} u^{(\nu_1/2)-1} \Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{2^{(\nu_1+\nu_2)/2}\Gamma(\nu_1/2)\Gamma(\nu_2/2)\left[\frac{1}{2}\left(1+\frac{\nu_1 u}{\nu_2}\right)\right]^{(\nu_1+\nu_2)/2}}$$

$$= \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \nu_1^{\nu_1/2}\nu_2^{\nu_2/2} u^{(\nu_1/2)-1}(\nu_2+\nu_1 u)^{-(\nu_1+\nu_2)/2}$$

si $u > 0$ y 0 si u $\leqq 0$, el cual es el resultado requerido.

4.46. Demuestre que la distribución F es unimodal en el valor $\left(\frac{\nu_1-2}{\nu_1}\right)\left(\frac{\nu_2}{\nu_2+2}\right)$ si $\nu_1 > 2$.

La moda se ubica en el valor máximo de la función de densidad. A diferencia de una constante, la función de densidad de la distribución F es

$$u^{(\nu_1/2)-1}(\nu_2+\nu_1 u)^{-(\nu_1+\nu_2)/2}$$

Si esto tiene un máximo relativo, ocurrirá si la derivada es cero, es decir,

$$\left(\frac{\nu_1}{2}-1\right)u^{(\nu_1/2)-2}(\nu_2+\nu_1 u)^{-(\nu_1+\nu_2)/2} - u^{(\nu_1/2)-1}\nu_1\left(\frac{\nu_1+\nu_2}{2}\right)(\nu_2+\nu_1 u)^{-[(\nu_1+\nu_2)/2]-1} = 0$$

Dividiendo por $u^{(\nu_1/2)-2}(\nu_2+\nu_1 u)^{-[(\nu_1+\nu_2)/2]-1}$, $u \neq 0$, encontramos

$$\left(\frac{\nu_1}{2}-1\right)(\nu_2+\nu_1 u) - u\nu_1\left(\frac{\nu_1+\nu_2}{2}\right) = 0 \quad \text{o} \quad u = \left(\frac{\nu_1-2}{\nu_1}\right)\left(\frac{\nu_2}{\nu_2+2}\right)$$

Usando la prueba de la segunda derivada podemos demostrar que esto es en realidad el máximo.

4.47. Usando la tabla de la distribución F del apéndice F, encuentre (*a*) $F_{0.95,10.15}$, (*b*) $F_{0.99,15.9}$, (*c*) $F_{0.05,8.30}$, (*d*) $F_{0.01,15.9}$.

(*a*) Del apéndice F, donde $\nu_1 = 10$, $\nu_2 = 15$, encontramos $F_{0.95,10.15} = 2.54$.

(*b*) Del apéndice F, donde $\nu_1 = 15$, $\nu_2 = 9$, encontramos $F_{0.99,15.9} = 4.96$.

(*c*) Según el teorema 4-8, página 126, $F_{0.05,8.30} = \dfrac{1}{F_{0.95,30,8}} = \dfrac{1}{3.08} = 0.325$.

(*d*) Según el teorema 4-8, página 126, $F_{0.01,15.9} = \dfrac{1}{F_{0.99,9,15}} = \dfrac{1}{3.89} = 0.257$.

RELACIONES ENTRE LAS DISTRIBUCIONES F, χ^2 y t

4.48. Verifique que (*a*) $F_{0.95} = t_{0.975}{}^2$, (*b*) $F_{0.99} = t_{0.995}{}^2$.

(*a*) Compare las entradas en la primera columna de la tabla $F_{0.95}$ en el apéndice F con aquéllas en la distribución t bajo $t_{0.975}$. Vemos que

$$161 = (12.71)^2, \quad 18.5 = (4.30)^2, \quad 10.1 = (3.18)^2, \quad 7.71 = (2.78)^2, \quad \text{etc.}$$

(*b*) Compare las entradas en la primera columna de la tabla $F_{0.99}$ en el apéndice F con aquéllas en la distribución t bajo $t_{0.995}$. Vemos que

$$4050 = (63.66)^2, \quad 98.5 = (9.92)^2, \quad 34.1 = (5.84)^2, \quad 21.2 = (4.60)^2, \quad \text{etc.}$$

4.49. Demuestre el teorema 4-9, página 127, el cual puede plantearse brevemente como

$$F_{1-p} = t^2_{1-(p/2)}$$

y por consiguiente, generalizar los resultados del problema 4.48.

Sean $v_1 = 1$, $v_2 = v$ en la función de densidad para la distribución F [*(45)*, página 126]. Entonces,

$$f(u) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{\nu}{2}\right)} \nu^{\nu/2} u^{-1/2} (\nu + u)^{-(\nu+1)/2}$$

$$= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \nu^{\nu/2} u^{-1/2} \nu^{-(\nu+1)/2} \left(1 + \frac{u}{\nu}\right)^{-(\nu+1)/2}$$

$$= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} u^{-1/2} \left(1 + \frac{u}{\nu}\right)^{-(\nu+1)/2}$$

para $u > 0$, y $f(u) = 0$ para $u \leqq 0$. Ahora, por definición de un valor percentil, F_{1-p} es el número tal que $P(U \leqq F_{1-p}) = 1 - p$. Por consiguiente,

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \int_0^{F_{1-p}} u^{-1/2} \left(1 + \frac{u}{\nu}\right)^{-(\nu+1)/2} du = 1 - p$$

En la integral, haga el cambio de la variable $t = +\sqrt{u}$:

$$2\,\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \int_0^{+\sqrt{F_{1-p}}} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2} dt = 1 - p$$

Comparando con *(42)*, página 125, vemos que el lado izquierdo de la última ecuación es igual a

$$2 \cdot P(0 < T \leqq +\sqrt{F_{1-p}})$$

donde T es la variable aleatoria que tiene distribución t de Student con v grados de libertad. Por consiguiente,

$$\frac{1-p}{2} = P(0 < T \leqq +\sqrt{F_{1-p}})$$

$$= P(T \leqq +\sqrt{F_{1-p}}) - P(T \leqq 0)$$

$$= P(T \leqq +\sqrt{F_{1-p}}) - \tfrac{1}{2}$$

donde usamos la simetría de la distribución t. Resolviendo, tenemos

$$P(T \leqq +\sqrt{F_{1-p}}) = 1 - \frac{p}{2}$$

Pero, por definición, $t_{1-(p/2)}$ es el número tal que

$$P(T \leqq t_{1-(p/2)}) = 1 - \frac{p}{2}$$

y este número se determina de manera única, dado que la función de densidad de la distribución t es estrictamente positivo. Por consiguiente,

$$+\sqrt{F_{1-p}} = t_{1-(p/2)} \qquad \text{o} \qquad F_{1-p} = t^2_{1-(p/2)}$$

lo cual queríamos demostrar.

4.50. Verifique el teorema 4-10, página 127, para (*a*) $p = 0.95$, (*b*) $p = 0.99$.

(*a*) Compare las entradas en la última fila de la tabla $F_{0.95}$ del apéndice F (correspondiente a $\nu_2 = \infty$) con las entradas bajo $\chi^2_{0.95}$ en el apéndice E. Entonces vemos que

$$3.84 = \frac{3.84}{1}, \qquad 3.00 = \frac{5.99}{2}, \qquad 2.60 = \frac{7.81}{3}, \qquad 2.37 = \frac{9.49}{4}, \qquad 2.21 = \frac{11.1}{5},$$

lo cual provee la verificación requerida.

(*b*) Compare las entradas en la última fila de la tabla $F_{0.99}$ del apéndice F (correspondiente a $\nu_2 = \infty$) con las entradas bajo $\chi^2_{0.95}$ en el apéndice E. Entonces vemos que

$$6.63 = \frac{6.63}{1}, \qquad 4.61 = \frac{9.21}{2}, \qquad 3.78 = \frac{11.3}{3}, \qquad 3.32 = \frac{13.3}{4}, \qquad 3.02 = \frac{15.1}{5},$$

lo cual provee la verificación requerida.

La demostración general del teorema 4-10 se deduce remplazando $\nu_2 \to \infty$ en la distribución F de la página 127.

LA DISTRIBUCIÓN NORMAL BIVARIADA

4.51. Supongamos que X y Y son variables aleatorias cuya función de densidad es la distribución normal bivariada. Demuestre que X y Y son independientes si y sólo si su coeficiente de correlación es cero.

Si el coeficiente de correlación $\rho = 0$, entonces la función de densidad normal bivariada (*49*), página 127, se convierte en

$$f(x, y) = \left[\frac{1}{\sigma_1\sqrt{2\pi}} e^{-(x-\mu_1)^2/2\sigma_1^2}\right]\left[\frac{1}{\sigma_2\sqrt{2\pi}} e^{-(y-\mu_2)^2/2\sigma_2^2}\right]$$

y dado que es un producto de la función de x sola y de y sola para todos los valores de x y y, se deduce que X y Y son independientes.

De manera inversa, si X y Y son independientes, $f(x, y)$ dado en (*49*) debe ser para todos los valores de x y y el producto de una función de x sola y de una función de y sola. Esto es posible sólo si $\rho = 0$.

DIVERSAS DISTRIBUCIONES

4.52. Encuentre la probabilidad de que en lanzamientos sucesivos de un dado balanceado, aparezca un 3 por vez primera en el quinto lanzamiento.

Método 1.

La probabilidad de no obtener 3 en el primer lanzamiento es $5/6$. De igual manera, la probabilidad de no obtener 3 en el segundo lanzamiento es $5/6$, etc. Entonces, la probabilidad de no obtener un 3 en los primeros cuatro lanzamientos es $(5/6)(5/6)(5/6)(5/6) = (5/6)^4$. Por consiguiente, dado que la probabilidad de obtener 3 en el quinto lanzamiento es de $1/6$, la probabilidad requerida es

$$\left(\frac{5}{6}\right)^4\left(\frac{1}{6}\right) = \frac{625}{7776}$$

Método 2 (usando fórmula).

Usando la distribución geométrica, página 127, con $p = 1/6$, $q = 5/6$, $x = 5$, vemos que la probabilidad requerida es

$$\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^4 = \frac{625}{7776}$$

4.53. Verifique la expresión dada para (*a*) la media, (*b*) la varianza de la distribución Weibull, página 128.

(*a*)
$$\mu = E(X) = \int_0^\infty abx^b e^{-ax^b}\,dx$$
$$= \frac{ab}{a^{1/b}} \int_0^\infty \left(\frac{u}{a}\right) e^{-u}\,\frac{1}{b}\,u^{(1/b)-1}\,du$$
$$= a^{-1/b} \int_0^\infty u^{1/b} e^{-u}\,du$$
$$= a^{-1/b}\Gamma\left(1+\frac{1}{b}\right)$$

donde sustituimos $u = ax^b$ para evaluar la integral.

(*b*)
$$E(X^2) = \int_0^\infty abx^{b+1} e^{-ax^b}\,dx$$
$$= \frac{ab}{a^{1/b}} \int_0^\infty \left(\frac{u}{a}\right)^{1+(1/b)} e^{-u}\,\frac{1}{b}\,u^{(1/b)-1}\,du$$
$$= a^{-2/b} \int_0^\infty u^{2/b} e^{-u}\,du$$
$$= a^{-2/b}\Gamma\left(1+\frac{2}{b}\right)$$

Entonces
$$\sigma^2 = E[(X-\mu)^2] = E(X^2) - \mu^2$$
$$= a^{-2/b}\left[\Gamma\left(1+\frac{2}{b}\right) - \Gamma^2\left(1+\frac{1}{b}\right)\right]$$

DIVERSOS PROBLEMAS

4.54. La probabilidad de que se gradúe un estudiante que está entrando a la universidad es 0.4. Determine la probabilidad de que de un grupo de 5 estudiantes se gradúe (*a*) ninguno, (*b*) 1, (*c*) al menos 1.

(*a*) P(ninguno se gradúe) $= {}_5C_0(0.4)^0(0.6)^5 = 0.07776$, o alrededor de 0.08.

(*b*) P(1 se gradúe) $= {}_5C_1(0.4)^1(0.6)^4 = 0.2592$, o alrededor de 0.26.

(*c*) P(al menos 1 se gradúe) $= 1 - P$(ninguno se gradúe) $= 0.92224$,o alrededor de 0.92.

4.55. ¿Cuál es la probabilidad de obtener un total de 9, en 6 lanzamientos de un par de dados, (*a*) dos veces, (*b*) al menos dos veces?

Cada una de las 6 maneras como puede caer el primer dado, se puede asociar con cada una de las otras 6 maneras como puede caer el segundo dado, de manera que hay $6 \cdot 6 = 36$ maneras como pueden caer ambos dados. Éstas son 1 en el primer dado y 1 en el segundo dado, 1 en el primer dado y 2 en el segundo dado, etc., denotado así: (1, 1), (1, 2), etc.

De estas 36 maneras, todas igualmente probables, si los dados son balanceados, ocurre un total de 9 en 4 casos: (3, 6), (4, 5), (5, 4), (6, 3). Entonces, la probabilidad de un total de 9 en un lanzamiento sencillo de un par de dados es p $= 4/36 = 1/9$, y la probabilidad de no obtener un total de 9 en un lanzamiento sencillo de un par de dados es $q = 1 - p = 8/9$.

(*a*) P(2 nueves en 6 lanzamientos) $= {}_6C_2\left(\frac{1}{9}\right)^2\left(\frac{8}{9}\right)^{6-2} = \frac{61{,}440}{531{,}441}$

(*b*) P(al menos 2 nueves) $= P$(2 nueves) + P(3 nueves) + P(4 nueves) + P(5 nueves) + P(6 nueves)
$$= {}_6C_2\left(\frac{1}{9}\right)^2\left(\frac{8}{9}\right)^4 + {}_6C_3\left(\frac{1}{9}\right)^3\left(\frac{8}{9}\right)^3 + {}_6C_4\left(\frac{1}{9}\right)^4\left(\frac{8}{9}\right)^2 + {}_6C_5\left(\frac{1}{9}\right)^5\frac{8}{9} + {}_6C_6\left(\frac{1}{9}\right)^2$$
$$= \frac{61{,}440}{531{,}441} + \frac{10{,}240}{531{,}441} + \frac{960}{531{,}441} + \frac{48}{531{,}441} + \frac{1}{531{,}441} = \frac{72{,}689}{531{,}441}$$

Otro método.

$$P(\text{al menos 2 nueves}) = 1 - P(0 \text{ nueves}) - P(1 \text{ nueve})$$

$$= 1 - {}_6C_0\left(\frac{1}{9}\right)^0\left(\frac{8}{9}\right)^6 - {}_6C_1\left(\frac{1}{9}\right)^1\left(\frac{8}{9}\right)^5 = \frac{72{,}689}{531{,}441}$$

4.56. Si la probabilidad de un tornillo defectuoso es 0.1, encuentre (*a*) la media, (*b*) la desviación estándar para la distribución de los tornillos defectuosos en un total de 400.

(*a*) Media $= np = 400(0.1) = 40$, es decir, podemos *esperar* que 40 tornillos sean defectuosos.

(*b*) Varianza $= npq = 400(0.1)(0.9) = 36$. De ahí, la desviación estándar $= \sqrt{36} = 6$.

4.57. Encuentre los coeficientes de (*a*) sesgo, (*b*) curtosis de la distribución del problema 4.56.

(*a*) $$\text{Coeficiente de sesgo} = \frac{q-p}{\sqrt{npq}} = \frac{0.9-0.1}{6} = 0.133$$

Dado que es positivo, la distribución está sesgada hacia la derecha.

(*b*) $$\text{Coeficiente de curtosis} = 3 + \frac{1-6pq}{npq} = 3 + \frac{1-6(0.1)(0.9)}{36} = 3.01$$

La distribución es ligeramente más empinada que la distribución normal.

4.58. Las notas de una evaluación corta de biología, que tuvo 10 preguntas, fueron 0, 1, 2, ..., 10 puntos, según el número de respuestas correctas. La nota promedio fue 6,7, y la desviación estándar 1,2. Suponiendo que las notas tienen distribución normal, determine (*a*) el porcentaje de estudiantes que obtuvieron 6 puntos, (*b*) la nota máxima del 10% de la clase que obtuvo la peor calificación, (*c*) la nota mínima del 10% de la clase que obtuvo la mejor calificación.

(*a*) Para aplicar la distribución normal a datos discretos, es necesario tratar los datos como si fueran continuos. Así, una nota de 6 puntos se considera como una nota de 5.5 a 6.5 puntos (*véase* figura 4-20).

$$5.5 \text{ unidades estándar} = (5.5 - 6.7)/1.2 = -1.0$$

$$6.5 \text{ unidades estándar} = (6.5 - 6.7)/1.2 = -0.17$$

$$\text{Proporción requerida} = \text{área entre } z = -1 \text{ y } z = -0.17$$

$$= (\text{área entre } z = -1 \text{ y } z = 0) - (\text{área entre } z = -0.17 \text{ y } z = 0)$$

$$= 0.3413 - 0.0675 = 0.2738 = 27\%$$

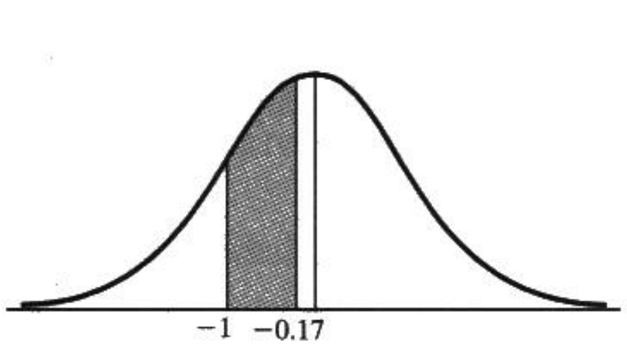


Fig. 4-20

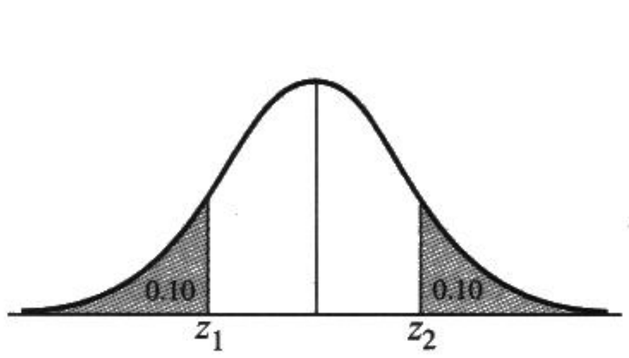


Fig. 4-21

(*b*) Sea x_1 el grado máximo requerido y z_1 su equivalente en unidades estándar. A partir de la figura 4-21, el área hacia la izquierda de z_1 es 10% = 0.10; de ahí,

$$\text{Área entre } z_1 \text{ y } 0 = 0.40$$

y $z_1 = -1.28$ (muy cercano)

Entonces, $z_1 = (x_1 - 6.7)/1.2 = 1.28$ y $x_1 = 5.2$ ó 5, aproximando al entero más cercano.

(*c*) Sea x_2 el grado mínimo requerido y z_2 su equivalente en unidades estándar. A partir de (*b*), por simetría, $z_2 = 1.28$. Entonces, $(x_2 - 6.7)/\ 1.2 = 1.28$, y $x = 8.2$ u 8, aproximando al entero más cercano.

4.59. Se usa un contador Geiger para contar la llegada de partículas radiactivas. Encuentre la probabilidad de que no se cuenten partículas en el tiempo t.

La figura 4-22 representa el eje x del tiempo con origen 0. La probabilidad de que se cuente una partícula en un tiempo pequeño Δt es proporcional a Δt y puede escribirse como $\lambda\Delta t$. Por consiguiente, la probabilidad de no conteos en un tiempo Δt es $1 - \lambda\Delta t$. Más precisamente, habrá términos adicionales que incluyen $(\Delta t)^2$ y órdenes mayores, pero son despreciables si Δt es pequeño.

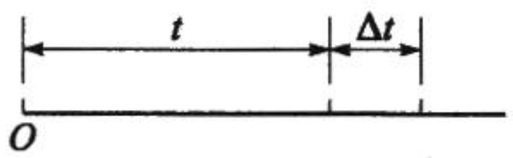


Fig. 4-22

Sea $P_0(t)$ la probabilidad de no conteos en el tiempo. Entonces, $P_0(t + \Delta t)$ es la probabilidad de no conteos en el tiempo $t + \Delta t$. Si las llegadas de las partículas se suponen como eventos independientes, la probabilidad de no conteo en el tiempo $t + \Delta t$ es el producto de la probabilidad de no conteo en el tiempo t y la probabilidad de no conteo en el tiempo Δt. Por consiguiente, los términos despreciables que involucran $(\Delta t)^2$ y mayores, tenemos

(*1*) $$P_0(t + \Delta t) = P_0(t)[1 - \lambda\Delta t]$$

A partir de (*1*) obtenemos

(*2*) $$\lim_{\Delta t \to 0} \frac{P_0(t + \Delta t) - P_0(t)}{\Delta t} = -\lambda P_0(t)$$

es decir,

(*3*) $$\frac{dP_0}{dt} = -\lambda P_0 \qquad \text{o} \qquad \frac{dP_0}{P_0} = -\lambda\, dt$$

Resolviendo (*3*) por integración obtenemos

$$\ln P_0 = -\lambda t + c_1 \qquad \text{o} \qquad P_0(t) = ce^{-\lambda t}$$

para determinar c, observe que si $t = 0$, $P_0(0) = c$ es la probabilidad requerida de no conteos en el tiempo cero, lo cual es, por supuesto, 1. Así $c = 1$ y la probabilidad requerida es

(*4*) $$P_0(t) = e^{-\lambda t}$$

4.60. En relación con el problema 4.59, encuentre la probabilidad de exactamente un conteo (que llegue una partícula) en el tiempo t.

Sea $P_1(t)$ la probabilidad de 1 conteo en el tiempo t, de manera que $P_1(t + \Delta t)$ es la probabilidad de 1 conteo en el tiempo $t + \Delta t$. Ahora, tendremos 1 conteo en el tiempo $t + \Delta t$ en los siguientes dos casos mutuamente excluyentes:

(i) 1 conteo en el tiempo t y 0 conteos en el tiempo Δt.

(ii) 0 conteos en el tiempo t y 1 conteo en el tiempo Δt.

La probabilidad de (i) es $p_1(t)(1 - \lambda\Delta t)$.

La probabilidad de (ii) es $P_0(t)\lambda\Delta t$.

Así, sin tener en cuenta los términos que incluyen $(\Delta t)^2$ y mayores,

(*1*) $$P_1(t + \Delta t) = P_1(t)(1 - \lambda\Delta t) + P_0(t)\lambda\Delta t$$

Esto puede escribirse

(*2*) $$\frac{P_1(t + \Delta t) - P_1(t)}{\Delta t} = \lambda P_0(t) - \lambda P_1(t)$$

Tomando el límite como $\Delta t \to 0$ y usando la expresión para $P_0(t)$ que se obtuvo en el problema 4.59, esto se convierte en

$$\frac{dP_1}{dt} = \lambda e^{-\lambda t} - \lambda P_1 \tag{3}$$

o

$$\frac{dP_1}{dt} + \lambda P_1 = \lambda e^{-\lambda t} \tag{4}$$

Multiplicando por $e^{\lambda t}$, puede escribirse,

$$\frac{d}{dt}(e^{\lambda t} P_1) = \lambda \tag{5}$$

de lo cual se obtiene integrando

$$P_1(t) = \lambda t e^{-\lambda t} + c_2 e^{-\lambda t} \tag{6}$$

Si $t = 0$, $P_1(0)$ es la probabilidad de 1 conteo en el tiempo 0, el cual es cero. Usando esto en (6), encontramos $c_2 = 0$. Por consiguiente,

$$P_1(t) = \lambda t e^{-\lambda t} \tag{7}$$

Si continuamos de esta manera, podemos demostrar que la probabilidad de exactamente n conteos en el tiempo t está dada por

$$P_n(t) = \frac{(\lambda t)^n e^{-\lambda t}}{n!} \tag{8}$$

la cual es la distribución de Poisson.

Problemas suplementarios

LA DISTRIBUCIÓN BINOMIAL

4.61. Encuentre la probabilidad de que lanzando 6 veces una moneda balanceada aparezcan (*a*) 0, (*b*) 1, (*c*) 2, (*d*) 3, (*e*) 4, (*f*) 5, (*g*) 6 caras.

4.62. Encuentre la probabilidad de (*a*) 2 o más caras, (*b*) menos de 4 caras, en un lanzamiento sencillo de 6 monedas balanceadas.

4.63. Si X es el número de caras en un lanzamiento sencillo de 4 monedas balanceadas, encuentre (*a*) $P(X = 3)$, (*b*) $P(X < 2)$, (*c*) $P(X \leqq 2)$, (*d*) $P(1 < X \leqq 3)$.

4.64. De 800 familias, cada una con 5 hijos, ¿cuántas esperaría usted que tuvieran (*a*) 3 niños, (*b*) 5 niñas, (*c*) bien sea 2 ó 3 niños? Suponga probabilidades iguales tanto para niños como para niñas.

4.65. Encuentre la probabilidad de obtener un total de 11 (*a*) una vez, (*b*) dos veces, en dos lanzamientos de un par de dados balanceados.

4.66. ¿Cuál es la probabilidad de obtener exactamente un 9 una vez en 3 lanzamientos de un par de dados?

4.67. Encuentre la probabilidad de acertar correctamente al menos 6 de 10 respuestas en un examen de falso y verdadero.

4.68. Un vendedor de seguros vende pólizas a 5 hombres, todos con la misma edad e idénticas condiciones de buena salud. Según las tablas de seguros, la probabilidad de que un hombre de esta edad en particular esté vivo dentro de 30 años es 2/3. Encuentre la probabilidad de que en 30 años estén vivos (*a*) los 5 hombres, (*b*) al menos 3 hombres, (*c*) sólo 2 hombres, (*d*) al menos 1 hombre.

4.69. Calcule (*a*) la media, (*b*) la desviación estándar, (*c*) el coeficiente de sesgo y (*d*) el coeficiente de curtosis para la distribución binomial en la cual $p = 0.7$ y $n = 60$. Interprete los resultados.

4.70. Demuestre que si la distribución binomial con $n = 100$ es simétrica, su coeficiente de curtosis es 2.9.

4.71. Para la distribución binomial calcule (*a*) $\sum (x-\mu)^3 f(x)$, (*b*) $\sum (x-\mu)^4 f(x)$.

LA DISTRIBUCIÓN NORMAL

4.72. En una evaluación de estadística la media fue 78 y la desviación estándar, 10. (*a*) Determine los puntajes estándar de dos estudiantes cuyas notas fueron 93 y 62, respectivamente. (*b*) Determine las notas de dos estudiantes cuyos puntajes estándar fueron -0.6 y 1.2, respectivamente.

4.73. Encuentre (*a*) la media, (*b*) la desviación estándar de una evaluación en la que las notas de 70 y 88 corresponden a puntajes estándar de -0.6 y 1.4, respectivamente.

4.74. Encuentre el área bajo la curva normal entre (*a*) $z = -1.20$ y $z = 2.40$, (*b*) $z = 1.23$ y $z = 1.87$, (*c*) $z = -2.35$ y $z = -0.50$.

4.75. Encuentre el área bajo la curva normal (*a*) hacia la izquierda de $z = -1.78$, (*b*) hacia la izquierda de $z = 0.56$, (*c*) a la derecha de $z = -1.45$, (*d*) correspondiente a $z \geqq 2.16$, (*e*) correspondiente a $-0.80 \leqq z \leqq 1.53$, (*f*) a la izquierda de $z = -2.52$ y a la derecha de $z = 1.83$.

4.76. Si Z tiene distribución normal con media 0 y varianza 1, encuentre: (*a*) $P(Z \geqq -1.64)$, (*b*) $P(-1.96 \leqq Z \leqq 1.96)$, (*c*) $P(|Z| \geqq 1)$.

4.77. Encuentre los valores de z tales que (*a*) el área a la derecha de z sea 0.2266, (*b*) el área a la izquierda de z sea 0.0314, (*c*) el área entre -0.23 y z sea 0.5722, (*d*) el área entre 1.15 y z sea 0.0730, (*e*) el área entre $-z$ y z sea 0.9000.

4.78. Encuentre z_1 si $P(Z \geqq z_1) = 0.84$, donde z tiene distribución normal con media 0 y varianza 1.

4.79. Si X tiene distribución normal con media 5 y desviación estándar 2, encuentre $P(X > 8)$.

4.80. Si el largo de 300 varillas tiene distribución normal con media 68.0 centímetros y desviación estándar 3.0 centímetros, ¿cuántas varillas tendrán largo (*a*) mayor de 72 centímetros, (*b*) menor o igual a 64 centímetros, (*c*) entre 65 y 71 centímetros, inclusive, (*d*) igual a 68 centímetros? Suponga que las medidas se anotaron aproximando al centímetro más cercano.

4.81. Si los diámetros de ciertas balineras tienen distribución normal con media 0.6140 centímetros y desviación estándar 0.0025 centímetros, determine el porcentaje de balineras con diámetros (*a*) entre 0.610 y 0.618 centímetros inclusive, (*b*) mayores que 0.617 centímetros, (*c*) menores que 0.608, (*d*) iguales que 0.615 centímetros.

4.82. La nota promedio en un examen final fue 72 y la desviación estándar fue 9. El 10% de los mejores estudiantes recibirán A. ¿Cuál es la nota mínima necesaria para que un estudiante obtenga una A?

4.83. Si un conjunto de medidas tiene distribución normal, ¿qué porcentaje de ellas diferirá de la media por (*a*) más de la mitad de la desviación estándar, (*b*) menos de tres cuartos de la desviación estándar?

4.84. Si μ es la media y σ es la desviación estándar de un conjunto de medidas que tienen distribución normal, ¿qué porcentaje de las medidas (*a*) están dentro del rango $\mu \pm 2\sigma$, (*b*) están por fuera del rango $\mu \pm 1.2\sigma$, (*c*) son mayores que $\mu - 1.5\sigma$?

4.85. Encuentre la constante a del problema 4.84 tal que el porcentaje de los casos (*a*) dentro del rango $\mu \pm a\sigma$ sea 75%, (*b*) menor que $\mu - a\sigma$ sea 22%.

APROXIMACIÓN NORMAL A LA DISTRIBUCIÓN BINOMIAL

4.86. Encuentre la probabilidad de que en 200 lanzamientos de una moneda haya (*a*) entre 80 y 120 caras inclusive, (*b*) menos de 90 caras, (*c*) menos de 85 o más de 115 caras, (*d*) exactamente 100 caras.

4.87. Encuentre la probabilidad de que en un examen de falso y verdadero un estudiante acierte (*a*) 12 o más respuestas de 20 (*b*) 24 o más respuestas de 40.

4.88. Una máquina fabrica tornillos de los cuales 10% son defectuosos. Encuentre la probabilidad de que en una muestra tomada al azar de 400 tornillos producidos por esta máquina (*a*) máximo 30, (*b*) entre 30 y 50, (*c*) entre 35 y 45, (*d*) 65 o más, de los tornillos sean defectuosos.

4.89. Encuentre la probabilidad de obtener más de 25 sietes en 100 lanzamientos de un par de dados balanceados.

LA DISTRIBUCIÓN DE POISSON

4.90. Si el 3% de las bombillas eléctricas que fabrica una compañía son defectuosas, encuentre la probabilidad de que en una muestra de 100 bombillas, (*a*) 0, (*b*) 1, (*c*) 2, (*d*) 3, (*e*) 4, (*f*) 5, sean defectuosas.

4.91. En el problema 4.90 encuentre la probabilidad de que (*a*) más de 5, (*b*) entre 1 y 3, (*c*) menos o 2, bombillas sean defectuosas.

4.92. Una bolsa contiene 1 canica roja y 7 blancas. Se saca una canica de la bolsa y se observa su color. Luego, la canica se devuelve a la bolsa y se mezcla muy bien el contenido. Encuentre la probabilidad de que en 8 retiros de la manera mencionada se seleccione la canica roja exactamente 3 veces, (*a*) usando la distribución binomial, (*b*) usando la aproximación de Poisson a la distribución binomial.

4.93. Según la oficina nacional de estadística vital del departamento de salud y servicios humanos de Estados Unidos, el promedio total de ahogamientos accidentales al año en Estados Unidos es 3.0 por cada 100,000 habitantes. Encuentre la probabilidad de que en una ciudad de 200,000 habitantes haya (*a*) 0, (*b*) 2, (*c*) 6, (*d*) 8, (*e*) entre 4 y 8, (*f*) menos de 3, ahogados por accidente al año.

4.94. Demuestre que si X_1 y X_2 son variables de Poisson independientes con parámetros respectivos λ_1 y λ_2, entonces $X_1 + X_2$ tiene distribución de Poisson con parámetro $\lambda_1 + \lambda_2$. (Ayuda: use la función generadora de momento.) Generalice el resultado a n variables.

DISTRIBUCIÓN MULTINOMIAL

4.95. Se lanza 6 veces un dado balanceado. Encuentre la probabilidad de que (*a*) aparezcan 1 "uno", 2 "doses", 3 "treses", (*b*) cada lado aparezca una vez.

4.96. Una caja contiene un número muy grande de canicas rojas, blancas, azules y amarillas en las proporciones 4:3:2:1. Encuentre la probabilidad de que en 10 retiros se saquen (*a*) 4 canicas rojas, 3 blancas, 2 azules y 1 amarilla, (*b*) 8 canicas rojas y 2 amarillas.

4.97. Encuentre la probabilidad de no obtener 1, 2 ó 3 en 4 lanzamientos de un dado balanceado.

LA DISTRIBUCIÓN HIPERGEOMÉTRICA

4.98. Una caja contiene 5 canicas rojas y 10 blancas. Si se seleccionan al azar 8 canicas (sin remplazo), determine la probabilidad de que (*a*) 4 sean rojas, (*b*) todas sean blancas, (*c*) al menos 1 sea roja.

4.99. Si se van a escoger al azar 13 cartas (sin remplazo) de una baraja común de 52 naipes, encuentre la probabilidad de que (*a*) 6 sean figuras, (*b*) ninguna sea figura.

4.100. De 60 aspirantes a una universidad, 40 son del Este. Si se van a seleccionar al azar 20, encuentre la probabilidad de que (*a*) 10, (*b*) no más de 2, vengan del Este.

LA DISTRIBUCIÓN UNIFORME

4.101. Si X está distribuida uniformemente en $-2 \leqq x \leqq 2$, encuentre (*a*) $P(X < 1)$, (*b*) $P(|X - 1| \geqq 1/2)$.

4.102. Encuentre (*a*) el tercer, (*b*) el cuarto momento alrededor de la media de una distribución uniforme.

4.103. Determine el coeficiente de (*a*) sesgo, (*b*) curtosis de una distribución uniforme.

4.104. Si X y Y son independientes y ambas tienen distribución uniforme en el intervalo 0 a 1, encuentre $P(|X - Y| \geqq 1/2)$.

LA DISTRIBUCIÓN DE CAUCHY

4.105. Suponga que X tiene distribución de Cauchy según (*29*), página 123, con $a = 2$. Encuentre (*a*) $P(X < 2)$, (*b*) $P(X^2 \geqq 12)$.

4.106. Demuestre que si X_1 y X_2 son independientes y tienen la misma distribución de Cauchy, entonces sus medias aritméticas también tienen dicha distribución.

4.107. Si X_1 y X_2 son independientes y están distribuidas normalmente con media 0 y varianza 1, demuestre que $Y = X_1/X_2$ tiene distribución de Cauchy.

LA DISTRIBUCIÓN GAMMA

4.108. Una variable aleatoria X tiene distribución gamma con $\alpha = 3$, $\beta = 2$. Encuentre (*a*) $P(X \leqq 1)$, (*b*) $P(1 \leqq X \leqq 2)$.

LA DISTRIBUCIÓN CHI CUADRADO

4.109. Para una distribución chi cuadrado con 12 grados de libertad, encuentre el valor χ_c^2 tal que (*a*) el área a la derecha de χ_c^2 sea 0.05, (*b*) el área a la izquierda de χ_c^2 sea 0.99, (*c*) el área a la derecha de χ_c^2 sea 0.025.

4.110. Encuentre los valores de χ^2 para los cuales el área de la cola del lado derecho de la distribución χ^2 sea 0.05, si el número de grados de libertad v es igual a (*a*) 8, (*b*) 19, (*c*) 28, (*d*) 40.

4.111. Desarrolle el problema 4.110 si el área bajo la cola derecha es 0.01.

4.112. (*a*) Encuentre χ_1^2 y χ_2^2 tales que el área bajo la distribución χ^2 correspondiente a $v = 20$ entre χ_2^2 y χ_1^2 sea 0.95, suponiendo áreas iguales a la derecha de χ_2^2 e izquierda de χ_1^2. (*b*) Demuestre que si el supuesto de áreas iguales en la parte (*a*) no se hace, los valores χ_1^2 y χ_2^2 no son únicos.

4.113. Si la variable U tiene distribución chi cuadrado con $v = 7$, encuentre χ_1^2 y χ_2^2 tales que (*a*) $P(U > \chi_2^2) = 0.025$, (*b*) $P(U < \chi_1^2) = 0.50$, (*c*) $P(\chi_1^2 \leqq U \leqq \chi_2^2) = 0.90$.

4.114. Encuentre (*a*) $\chi_{0.05}^2$ y (*b*) $\chi_{0.95}^2$ para $\nu = 150$.

4.115. Encuentre (*a*) $\chi_{0.025}^2$ y (*b*) $\chi_{0.975}^2$ para $\nu = 250$.

DISTRIBUCIÓN *t* DE STUDENT

4.116. Para una distribución de Student con15 grados de libertad, encuentre el valor de t_1 tal que (*a*) el área a la derecha de t_1 sea 0.01, (*b*) el área a la izquierda de t_1 sea 0.95, (*c*) el área a la derecha de t sea 0.10, (*d*) el área combinada a la derecha de t_1 y a la izquierda de $-t_1$ sea 0.01, (*e*) el área entre $-t_1$ y t sea 0.95.

4.117. Encuentre los valores de t para los cuales el área de la cola derecha de la distribución t es 0.01, si el número de grados de libertad v es igual a (*a*) 4, (*b*) 12, (*c*) 25, (*d*) 60, (*e*) 150.

4.118. Encuentre los valores de t_1 para la distribución de Student que satisfaga cada una de las siguientes condiciones: (*a*) el área entre $-t_1$ y t_1 es 0.90 y $v = 25$, (*b*) el área a la izquierda de $-t_1$ es 0.025 y $v = 20$, (*c*) el área combinada a la derecha de t_1 y a la izquierda de $-t_1$ es 0.01 y $v = 5$, (*d*) el área a la derecha de t_1 es 0.55 y $v = 16$.

4.119. Si una variable U tiene distribución de Student con $v = 10$, encuentre la constante c tal que (*a*) $P(U > c) = 0.05$, (*b*) $P(-c \leqq U \leqq c) = 0.98$, (*c*) $P(U \leqq c) = 0.20$, (*d*) $P(U \geqq c) = 0.90$.

LA DISTRIBUCIÓN F

4.120. Calcule cada uno de los siguientes:
(*a*) $F_{0.95,15,12}$; (*b*) $F_{0.99,120,60}$; (*c*) $F_{0.99,60,24}$; (*d*) $F_{0.01,30,12}$; (*e*) $F_{0.05,9,20}$; (*f*) $F_{0.01,8,8}$.

Respuestas a los problemas suplementarios

4.61. (*a*) 1/64 (*b*) 3/32 (*c*) 15/64 (*d*) 5/16 (*e*) 15/64 (*f*) 3/32 (*g*) 1/64

4.62. (*a*) 57/64 (*b*) 21/32

4.63. (*a*) 1/4 (*b*) 5/16 (*c*) 11/16 (*d*) 5/8

4.64. (*a*) 250 (*b*) 25 (*c*) 500

4.65. (*a*) 17/162 (*b*) 1/324

4.66. 64/243

4.67. 193/512

4.68. (*a*) 32/243 (*b*) 192/243 (*c*) 40/243 (*d*) 242/243

4.69. (*a*) 42 (*b*) 3.550 (*c*) −0.1127 (*d*) 2.927

4.71. (*a*) $npq(q - p)$ (*b*) $npq(1 - 6pq) + 3n^2p^2q^2$

4.72. (*a*) 1.5, −1.6 (*b*) 72, 90

4.73. (*a*) 75.4 (*b*) 9

4.74. (*a*) 0.8767 (*b*) 0.0786 (*c*) 0.2991

4.75. (*a*) 0.0375 (*b*) 0.7123 (*c*) 0.9265 (*d*) 0.0154 (*e*) 0.7251 (*f*) 0.0395

4.76. (*a*) 0.9495 (*b*) 0.9500 (*c*) 0.6826

4.77. (*a*) 0.75 (*b*) −1.86 (*c*) 2.08 (*d*) 1.625 o 0.849 (*e*) ±1.645

4.78. −0.995

4.79. 0.0668

4.80. (*a*) 20 (*b*) 36 (*c*) 227 (*d*) 40

4.81. (*a*) 93% (*b*) 8.1% (*c*) 0.47% (*d*) 15%

4.82. 84

4.83. (*a*) 61.7% (*b*) 54.7%

4.84. (*a*) 95.4% (*b*) 23.0% (*c*) 93.3%

4.85. (*a*) 1.15 (*b*) 0.77

4.86. (*a*) 0.9962 (*b*) 0.0687 (*c*) 0.0286 (*d*) 0.0558

4.87. (*a*) 0.2511 (*b*) 0.1342

4.88. (*a*) 0.0567 (*b*) 0.9198 (*c*) 0.6404 (*d*) 0.0079

4.89. 0.0089

4.90. (*a*) 0.04979 (*b*) 0.1494 (*c*) 0.2241 (*d*) 0.2241 (*e*) 0.1680 (*f*) 0.1008

4.91. (*a*) 0.0838 (*b*) 0.5976 (*c*) 0.4232

4.92. (*a*) 0.05610 (*b*) 0.06131

4.93. (*a*) 0.00248 (*b*) 0.04462 (*c*) 0.1607 (*d*) 0.1033 (*e*) 0.6964 (*f*) 0.0620

4.95. (*a*) 5/3888 (*b*) 5/324

4.96. (*a*) 0.0348 (*b*) 0.000295

4.97. 1/16

4.98. (*a*) 70/429 (*b*) 1/143 (*c*) 142/143

4.99. (*a*) $\binom{13}{6}\binom{39}{7}\Big/\binom{52}{13}$ (*b*) $\binom{13}{0}\binom{39}{13}\Big/\binom{52}{13}$

4.100. (*a*) $\binom{40}{10}\binom{20}{10}\Big/\binom{60}{20}$ (*b*) $[({}_{40}C_0)({}_{20}C_{20}) + ({}_{40}C_1)({}_{20}C_{19}) + ({}_{40}C_2)({}_{20}C_{18})]/{}_{60}C_{20}$

4.101. (*a*) 3/4 (*b*) 3/4

4.102. (*a*) 0 (*b*) $(b-a)^4/80$

4.103. (*a*) 0 (*b*) 9/5

4.104. 1/4

4.105. (*a*) 3/4 (*b*) 1/3

4.108. (*a*) $1 - \dfrac{13}{8\sqrt{e}}$ (*b*) $\dfrac{13}{8}\,e^{-1/2} - \dfrac{5}{2}\,e^{-1}$

4.109. (*a*) 21.0 (*b*) 26.2 (*c*) 23.3

4.110. (*a*) 15.5 (*b*) 30.1 (*c*) 41.3 (*d*) 55.8

4.111. (*a*) 20.1 (*b*) 36.2 (*c*) 48.3 (*d*) 63.7

4.112. (*a*) 9.59 y 34.2

4.113. (*a*) 16.0 (*b*) 6.35 (*c*) suponiendo áreas iguales en las dos colas, $\chi_1^2 = 2.17$ y $\chi_2^2 = 14.1$

4.114. (*a*) 122.5 (*b*) 179.2

4.115. (*a*) 207.7 (*b*) 295.2

4.116. (*a*) 2.60 (*b*) 1.75 (*c*) 1.34 (*d*) 2.95 (*e*) 2.13

4.117. (*a*) 3.75 (*b*) 2.68 (*c*) 2.48 (*d*) 2.39 (*e*) 2.33

4.118. (*a*) 1.71 (*b*) 2.09 (*c*) 4.03 (*d*) −0.128

4.119. (*a*) 1.81 (*b*) 2.76 (*c*) −0.879 (*d*) −1.37

4.120. (*a*) 2.62 (*b*) 1.73 (*c*) 2.40 (*d*) 0.352 (*e*) 0.340 (*f*) 0.166

PARTE II

Estadística

CAPÍTULO 5

Teoría de muestreo

POBLACIÓN Y MUESTRA. INFERENCIA ESTADÍSTICA

En la práctica con frecuencia estamos interesados en establecer conclusiones válidas sobre un grupo grande de individuos u objetos. En lugar de examinar el grupo completo, llamado *población*, lo cual puede ser difícil o imposible, examinamos solamente una parte pequeña de esa población, la cual llamamos *muestra*. Hacemos esto con el ánimo de inferir ciertos hechos sobre la población a partir de los resultados que encontramos en la muestra, un proceso conocido como *inferencia estadística*. El proceso de obtención de muestras se denomina *muestreo*.

EJEMPLO 5.1. Podemos querer sacar conclusiones sobre las estaturas (o pesos) de 12,000 estudiantes adultos (la población) al examinar tan sólo 100 estudiantes (la muestra) seleccionados de esa población.

EJEMPLO 5.2. Podemos querer sacar conclusiones sobre el porcentaje de tornillos defectuosos que se producen en una fábrica durante cierta semana de 6 días si examinamos 20 tornillos diarios, obtenidos en diferentes momentos de cada día. En este caso, todos los tornillos que se producen durante la semana forman la población, mientras que los 120 tornillos seleccionados constituyen la muestra.

EJEMPLO 5.3. Podemos querer establecer conclusiones sobre la imparcialidad de una moneda en particular lanzándola repetidamente. La población consistirá en todos los posibles lanzamientos de la moneda. Se puede obtener una muestra examinando, digamos, los primeros 60 lanzamientos de la moneda y anotando el porcentaje de caras y sellos.

EJEMPLO 5.4. Podemos querer obtener conclusiones sobre los colores de 200 canicas (la población) en una urna seleccionando una muestra de 20 canicas de la urna, donde cada canica seleccionada se regresa luego de que se observa su color.

Debemos tener en cuenta varias cosas. Primero, la palabra *población* no necesariamente tiene el mismo significado que tiene en el lenguaje común, tal como decir que "la población" de una ciudad es de 180,000 personas. Segundo, la palabra *población* se usa con frecuencia para denominar las observaciones o medidas en lugar de los individuos u objetos. En el ejemplo 5.1 podemos hablar de la población de 12,000 estaturas (o pesos), mientras que en el ejemplo 5.4 podemos hablar de la población de los 200 colores en la urna (algunos de los cuales pueden ser los mismos). Tercero, la población puede ser finita o infinita, siendo su número el *tamaño de la población*, generalmente denominada como N. De manera similar, el número en la muestra se denota por *tamaño de la muestra* o *tamaño muestral*, se denota como n, y generalmente es finito. En el ejemplo 5.1, $N = 12{,}000$, $n = 100$, mientras que en el ejemplo 5.3, N es infinito y $n = 60$.

MUESTREO CON Y SIN REMPLAZO

Si retiramos un objeto de una urna, tenemos la opción de devolverlo o no antes de sacar el siguiente. En el primer caso, un objeto en particular puede salir una y otra vez, mientras que en el segundo, sólo puede salir una vez. Tomar muestras en donde cada miembro de la población puede escogerse más de una vez se llama *muestreo con remplazo*, mientras que tomar muestras en donde cada miembro no se puede escoger más de una vez se llama *muestreo sin remplazo*.

Una población finita de la que se toman muestras con remplazo puede considerarse teóricamente infinita dado que se pueden tomar muestras de cualquier tamaño sin agotar la población. Para propósitos más prácticos, tomar muestras de una población finita muy grande puede considerarse como tomar muestras de una población infinita.

MUESTRAS ALEATORIAS. NÚMEROS ALEATORIOS

Claramente, la confiabilidad de las conclusiones obtenidas concernientes a una población dependen de si la muestra se tomó adecuadamente, para que represente a la población lo suficientemente bien. Uno de los problemas importantes de la inferencia estadística es precisamente cómo obtener una muestra.

Una manera de hacer esto para poblaciones finitas es asegurar que cada miembro de la población tenga la misma probabilidad de estar en la muestra, lo cual se denomina *muestra aleatoria*. Una muestra aleatoria puede lograrse para muestras de poblaciones relativamente pequeñas extrayendo lotes, o de manera equivalente, usando una tabla de *números aleatorios* (apéndice H), especialmente construida para tales propósitos (*véase* problema 5.43).

Dado que la inferencia de muestra a población no puede ser segura, debemos usar el lenguaje de la probabilidad en cualquier afirmación de conclusiones.

PARÁMETROS DE LA POBLACIÓN

Se considera que una población es conocida cuando conocemos la distribución de probabilidad $f(x)$ (función de probabilidad o función de densidad) de la variable aleatoria asociada X. En el ejemplo 5.1 si X es una variable aleatoria cuyos valores son las estaturas (o pesos) de 12,000 estudiantes, entonces X tiene distribución de probabilidad $f(x)$.

Si, por ejemplo, X tiene distribución normal, decimos que la población tiene *distribución normal* o que tenemos una *población normal*. De igual manera, si X tiene distribución binomial, decimos que la población tiene *distribución binomial* o que tenemos una *población binomial*.

Habrá ciertas medidas que aparecen en $f(x)$, tales como μ y σ en el caso de la distribución normal, o p en el caso de la distribución binomial. Otras cantidades como la mediana, los momentos y el sesgo pueden determinarse en términos de éstos. Tales medidas se denominan con frecuencia *parámetros poblacionales*. Cuando tenemos una población dada, de manera que conocemos $f(x)$, entonces los parámetros de la población también se conocen.

Un problema importante surge cuando la distribución de probabilidad $f(x)$ de la población no se conoce precisamente, aunque podemos tener alguna idea, o al menos estar en capacidad de plantear alguna hipótesis concerniente al comportamiento general de $f(x)$. Por ejemplo, podemos tener alguna razón para suponer que una población en particular tiene distribución normal. En tal caso podemos no conocer uno o ambos valores μ y σ y entonces queremos establecer una inferencia estadística sobre ellos.

ESTADÍSTICO MUESTRAL

Podemos tomar muestras aleatorias de la población y luego usarlas con el fin de obtener valores que ayuden a estimar y a probar hipótesis sobre los parámetros de la población.

A manera de ilustración consideremos el ejemplo 5.1, donde X es una variable aleatoria cuyos valores son las diferentes estaturas. Para obtener una muestra de 100, primero debemos escoger de la población un individuo al azar. Éste puede tener cualquier valor, digamos x_1, de todas las estaturas posibles, y denominamos x_1 al

valor de una variable aleatoria X_1, donde se usa el subíndice 1 dado que corresponde al individuo que se seleccionó en primer lugar. De igual manera escogemos el segundo individuo para la muestra, el cual puede tener cualquiera de los valores x_2 de las estaturas posibles, y x_2 se toma como el valor de una variable aleatoria X_2. Continuamos con este proceso hasta llegar a X_{100} dado que el tamaño de la muestra es 100. Para hacer esta explicación más sencilla, supongamos que el muestreo es con remplazo, así que existe la posibilidad de escoger el mismo individuo más de una vez. En este caso, dado que el tamaño muestral es mucho más pequeño que el de la población, muestrear sin remplazo da prácticamente el mismo resultado que hacerlo con remplazo.

En el caso general, una muestra de tamaño n puede describirse por los valores $x_1, x_2, \ldots, x_n$ de las variables aleatorias $X_1, X_2, \ldots, X_n$. En el caso de tomar muestras con remplazo, $X_1, X_2, \ldots, X_n$ serán variables aleatorias independientes, idénticamente distribuidas con distribución de probabilidad $f(x)$. Su distribución conjunta será

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = f(x_1)f(x_2)\cdots f(x_n) \tag{1}$$

Cualquier medida que se obtenga a partir de una muestra tomada con el propósito de calcular un parámetro de la población se llama *estadístico muestral*, o simplemente *estadístico*. Matemáticamente, un estadístico muestral para una muestra de tamaño n puede definirse como la función de las variables aleatorias $X_1, \ldots, X_n$, es decir, $g(X_1, \ldots, X_n)$. La función $g(X_1, \ldots, X_n)$ es otra variable aleatoria, cuyos valores pueden representarse por $g(x_1, \ldots, x_n)$. La palabra *estadístico* se utiliza con frecuencia para la variable aleatoria o para sus valores, y su sentido particular se sobreentiende a partir del contexto en que se use.

En general, a cada parámetro de la población le corresponde un estadístico que se calcula a partir de la muestra. Usualmente, el método que se usa para obtener este estadístico a partir de la muestra es similar al que se utiliza para obtener el parámetro de una población finita, dado que la muestra consta de un conjunto finito de valores. Sin embargo, como podemos ver, esto no siempre produce el "mejor estimado", y uno de los problemas importantes de la teoría del muestreo es decidir cómo formar el estadístico muestral adecuado que dé el mejor estimado de un parámetro poblacional dado. Tales problemas se consideran en capítulos posteriores.

Donde sea posible trataremos de utilizar letras griegas, como μ y σ, para los valores de los parámetros poblacionales y, letras romanas, m, s, etc., para los valores de los estadísticos muestrales correspondientes.

DISTRIBUCIÓN DE LA MUESTRA

Como hemos visto, un estadístico muestral que se calcula a partir de $X_1, \ldots, X_n$ es una función de estas variables aleatorias y, por consiguiente, es en sí misma una variable aleatoria. La distribución de probabilidad de un estadístico muestral se llama *distribución muestral* del estadístico.

De manera alterna podemos considerar todas las muestras posibles de tamaño n que se pueden sacar de una población, y para cada muestra calculamos el estadístico. Así obtenemos la distribución del estadístico, la cual es la distribución muestral o distribución de la muestra.

Por supuesto, podemos calcular la media, la varianza, la desviación estándar y los momentos, entre otros, de una distribución muestral. La desviación estándar se denomina algunas veces *error estándar*.

LA MEDIA DE LA MUESTRA

Sean $X_1, X_2, \ldots, X_n$ las variables aleatorias independientes con distribución idéntica para una muestra aleatoria de tamaño n como se describió anteriormente. Entonces, la *media de la muestra* o *media muestral* es la variable aleatoria definida por

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \tag{2}$$

en analogía con (*3*), página 82. Si $x_1, x_2, \ldots, x_n$ denotan valores obtenidos en una muestra en particular de tamaño n, entonces la media para tal muestra se describe como

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{3}$$

EJEMPLO 5.5. Si una muestra de tamaño 5 da como resultado los valores muestrales 7, 9, 1, 6, 2, entonces la media muestral es

$$\bar{x} = \frac{7+9+1+6+2}{5} = 5$$

DISTRIBUCIÓN MUESTRAL DE MEDIAS

Sea $f(x)$ la distribución de probabilidad de cierta población dada y de la cual obtenemos una muestra de tamaño n. Entonces es natural que averigüemos la distribución de probabilidad del estadístico muestral $\bar{X}$, al cual llamamos *distribución muestral para la media de la muestra* o *distribución muestral de medias*. Los siguientes teoremas son importantes en relación con este concepto.

Teorema 5-1: La media de la distribución muestral de las medias, denotada como $\mu_{\bar{X}}$ está dada por

$$E(\bar{X}) = \mu_{\bar{X}} = \mu \tag{4}$$

donde μ es la media de la población.

El teorema 5-1 plantea que el valor esperado de la media muestral es la media de la población.

Teorema 5-2: Si una población es infinita y el muestreo es aleatorio o si la población es finita y el muestreo es con remplazo, entonces la varianza de la distribución muestral de medias, denotado por $\sigma^2_{\bar{X}}$, está dado por

$$E[(\bar{X} - \mu)^2] = \sigma^2_{\bar{X}} = \frac{\sigma^2}{n} \tag{5}$$

donde σ^2 es la varianza de la población.

Teorema 5-3: Si la población es de tamaño N, si el muestreo es sin remplazo y si el tamaño de la muestra es $n \leqq N$, entonces (5) se remplaza por

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \tag{6}$$

mientras que $\mu_{\bar{X}}$ está dado por (4).

Observe que (6) se reduce a (5) a medida que $N \to \infty$.

Teorema 5-4: Si la población de la que estamos tomando muestras tiene distribución normal con media μ y varianza σ^2, entonces la media muestral tiene distribución normal con media μ y varianza σ^2/n.

Teorema 5-5: Supongamos que la población de la que estamos tomando muestras tiene distribución de probabilidad con media μ y varianza σ^2, que no necesariamente es una distribución normal. Entonces la variable estandarizada asociada con $\bar{X}$, dada por

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{7}$$

es asintóticamente normal, es decir,

$$\lim_{n\to\infty} P(Z \leqq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2}\, du \tag{8}$$

El teorema 5-5 es una consecuencia del teorema del límite central, página 121. Aquí se supone que la población es infinita o que se toman muestras con remplazo. De otra manera, la expresión anterior es correcta si remplazamos $\sigma/\sqrt{n}$ en (7) por $\sigma_{\bar{X}}$ como está dado en (6).

DISTRIBUCIÓN MUESTRAL DE PROPORCIONES

Supongamos que una población es infinita y tiene distribución binomial con p y $q = 1 - p$ siendo las probabilidades respectivas de que cualquier miembro exhibe o no cierta propiedad. Por ejemplo, la población pueden ser todos los lanzamientos posibles de una moneda balanceada, en los cuales la probabilidad del evento caras es $p = \frac{1}{2}$.

Consideremos todas las muestras posibles de tamaño n sacadas de esta población, y para cada muestra determinemos el estadístico que es la proporción P de éxitos. En el caso de la moneda, P sería la proporción de caras que aparecen en n lanzamientos. Entonces obtenemos una distribución muestral de proporciones cuya media μ_P y desviación estándar σ_P están dadas por

$$\mu_P = p \qquad \sigma_P = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}} \tag{9}$$

lo cual puede obtenerse de (*4*) y (*5*) colocando $\mu = p$, $\sigma = \sqrt{pq}$.

Para valores grandes de n ($n \geqq 30$), la distribución muestral está muy cerca a una distribución normal, tal como se ve a partir del teorema 5-5.

Para poblaciones finitas, en las que el muestreo se hace sin remplazo, la segunda ecuación en (*9*) se reemplaza por $\sigma_{\bar{X}}$ como se da en (*6*) con $\sigma = \sqrt{pq}$.

Observe que las ecuaciones (*9*) se obtienen más fácilmente dividiendo por n la media y la desviación estándar (np y $\sqrt{npq}$) de la distribución binomial.

DISTRIBUCIÓN MUESTRAL DE DIFERENCIAS Y SUMAS

Supongamos que tenemos dos poblaciones. Para cada tamaño de muestra n_1 sacado de la primera población, calculemos un estadístico S_1. Esto da como resultado una distribución muestral para S_1 cuya media y desviación estándar escribimos como μ_{S_1} y σ_{S_1}, respectivamente. De igual manera, para cada muestra de tamaño n_2 sacada de la segunda población, calculemos un estadístico S_2, cuya media y desviación estándar son μ_2 y σ_{S2}, respectivamente.

Tomando todas las combinaciones posibles de estas muestras a partir de la población, obtenemos una distribución de diferencias, $S_1 - S_2$, la cual se denomina *distribución muestral de diferencias* de estadísticas. La media y la desviación estándar de esta distribución muestral, descrita por $\mu_{S_1 - S_2}$ y $\sigma_{S_1 - S_2}$, respectivamente, están dadas por

$$\mu_{S_1 - S_2} = \mu_{S_1} - \mu_{S_2} \qquad \sigma_{S_1 - S_2} = \sqrt{\sigma_{S_1}^2 + \sigma_{S_2}^2}, \tag{10}$$

siempre y cuando las muestras escogidas no dependan de ninguna manera una de la otra, es decir, que las muestras sean independientes (en otras palabras, que las variables aleatorias S_1 y S_2 sean independientes).

Si, por ejemplo, S_1 y S_2 son las medias muestrales de dos poblaciones, denominadas $\bar{X}_1$ y $\bar{X}_2$, respectivamente, entonces la distribución muestral de la diferencia de las medias para poblaciones infinitas con media y desviación estándar μ_1, σ_1 y μ_2, σ_2, respectivamente, está dada por

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2, \qquad \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{11}$$

usando (*4*) y (*5*). Este resultado también es válido para poblaciones finitas si el muestreo se hace con remplazo. La variable estandarizada

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \tag{12}$$

en ese caso está muy cerca de la distribución normal si n_1 y n_2 son grandes ($n_1, n_2 \geqq 30$). Se pueden obtener resultados similares para poblaciones finitas en las cuales el muestreo se hace sin remplazo, usando (*4*) y (*6*).

Se pueden obtener resultados correspondientes para distribuciones muestrales de diferencias de proporciones a partir de dos poblaciones con distribución binomial con parámetros p_1, q_1 y p_2, q_2, respectivamente. En este caso, S_1 y S_2 corresponden a la proporción de los éxitos P_1 y P_2, y las ecuaciones (*11*) dan como resultado

$$\mu_{P_1-P_2} = \mu_{P_1} - \mu_{P_2} = p_1 - p_2, \qquad \sigma_{P_1-P_2} = \sqrt{\sigma_{P_1}^2 + \sigma_{P_2}^2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \tag{13}$$

En lugar de tomar diferencias de estadísticos, algunas veces estamos interesados en la suma de ellos. En ese caso la distribución muestral de la suma de los estadísticos S_1 y S_2 tiene media y desviación estándar dados por

$$\mu_{S_1+S_2} = \mu_{S_1} + \mu_{S_2} \qquad \sigma_{S_1+S_2} = \sqrt{\sigma_{S_1}^2 + \sigma_{S_2}^2} \tag{14}$$

suponiendo que las muestras sean independientes. Se pueden obtener resultados similares a (*11*).

LA VARIANZA DE LA MUESTRA

Si $X_1, X_2, \ldots, X_n$ son las variables aleatorias de una muestra aleatoria de tamaño n, entonces la variable aleatoria que da la *varianza de la muestra* o *varianza muestral* está definida por analogía con (*14*), página 84, por

$$S^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n} \tag{15}$$

Ahora, en el teorema 5-1 encontramos que $E(\bar{X}) = \mu$, y sería bueno si también pudiéramos tener $E(S^2) = \sigma^2$. Siempre que el valor esperado de un estadístico sea igual al parámetro poblacional correspondiente, decimos que el estadístico es un *estimador insesgado* y que el valor es un *estimado insesgado* de este parámetro. Sin embargo (ver problema 5.20), resulta que

$$E(S^2) = \mu_{S^2} = \frac{n-1}{n}\sigma^2 \tag{16}$$

lo cual es muy cercano a σ^2 sólo para valores grandes de n (digamos, $n \geqq 30$). El estimador insesgado deseado se define por

$$\hat{S}^2 = \frac{n}{n-1}S^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1} \tag{17}$$

de manera que

$$E(\hat{S}^2) = \sigma^2 \tag{18}$$

Debido a esto, algunos estudiosos de la estadística prefieren definir la varianza muestral por $\hat{S}^2$ en vez de S^2 y simplemente remplazan n por $n - 1$ en el denominador de (*15*). Sin embargo, debemos continuar definiendo la varianza muestral según (*15*) porque al hacerlo, se simplifican muchos resultados posteriores.

EJEMPLO 5.6. En relación con el ejemplo 5.5, página 168, la varianza muestral tiene el valor de

$$s^2 = \frac{(4-6)^2 + (7-6)^2 + (5-6)^2 + (8-6)^2 + (6-6)^2}{5} = 2$$

mientras que un estimado insesgado está dado por

$$\hat{s}^2 = \frac{5}{4}s^2 = \frac{(4-6)^2 + (7-6)^2 + (5-6)^2 + (8-6)^2 + (6-6)^2}{4} = 2.5$$

Los resultados anteriores son válidos si el muestreo se hace de una población infinita o, de una población finita con remplazo. Si la toma de muestras se lleva a cabo sin remplazo de una población finita de tamaño N, entonces la media de la distribución muestral de varianzas está dada por

$$E(S^2) = \mu_{S^2} = \left(\frac{N-n}{N-1}\right)\left(\frac{n-1}{n\ n}\right)\sigma^2 \tag{19}$$

A medida que $N \to \infty$, esto se reduce a (*16*).

DISTRIBUCIÓN MUESTRAL DE VARIANZAS

Tomando todas las muestras aleatorias posibles de tamaño n, sacadas de una población y luego calculando la varianza para cada una, podemos obtener la distribución muestral de las varianzas. En lugar de encontrar la distribución muestral de S^2 o $\hat{S}^2$ en sí misma, conviene encontrar la distribución muestral de las variables aleatorias relacionadas.

$$\frac{nS^2}{\sigma^2} = \frac{(n-1)\hat{S}^2}{\sigma^2} = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{\sigma^2} \tag{20}$$

La distribución de esta variable aleatoria se describe en el siguiente teorema.

Teorema 5-6: Si se toman muestras aleatorias de tamaño n de una población que tiene distribución normal, entonces la variable de muestreo (*20*) tiene distribución chi cuadrado con $n - 1$ grados de libertad.

Debido al teorema 5-6, la variable en (*20*) se denota con frecuencia χ^2. Para una demostración de este teorema, *véase* el problema 5.22.

CASO EN QUE SE DESCONOCE LA VARIANZA DE LA POBLACIÓN

En los teoremas 5-4 y 5-5 encontramos que la variable estandarizada

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{21}$$

tiene distribución normal si la población de la que se tomaron muestras de tamaño n tiene distribución normal, mientras que es normal asintóticamente si la población no es normal siempre y cuando $n \geqq 30$. En (*21*) supusimos, por supuesto, que se conoce la varianza de la población σ^2.

Es natural preguntarse qué ocurriría si no conociéramos la varianza de la población. Una posibilidad es calcular la varianza de la población usando la varianza muestral y luego colocando la desviación estándar correspondiente en (*21*). Una mejor idea es remplazar la σ en (*21*) por la variable aleatoria $\hat{S}$ dando la desviación estándar muestral y buscar la distribución del estadístico correspondiente, el cual hemos designado por

$$T = \frac{\bar{X} - \mu}{\hat{S}/\sqrt{n}} = \frac{\bar{X} - \mu}{S/\sqrt{n-1}} \tag{22}$$

Después podemos demostrar, usando el teorema 4-6, página 126, que T tiene la distribución t de Student con $n - 1$ grados de libertad siempre que la variable aleatoria de la población tenga distribución normal. Planteamos esto en el siguiente teorema, el cual está demostrado en el problema 5.24.

Teorema 5-7: Si muestras aleatorias de tamaño n se toman de una población con distribución normal, entonces el estadístico (*22*) tiene distribución de Student con $n - 1$ grados de libertad.

DISTRIBUCIÓN MUESTRAL DE LAS RELACIONES DE VARIANZAS

En la página 169 indicamos cómo se pueden obtener distribuciones muestrales de diferencias, en especial de diferencias de medias. Usando la misma idea, podemos llegar a la distribución muestral de diferencias de varianzas, digamos $S_1^2 - S_2^2$. Sin embargo, resulta que esta distribución muestral es bastante complicada. En cambio, podemos considerar el estadístico S_1^2/S_2^2, ya que una proporción grande o pequeña indicará una diferencia grande mientras que una proporción casi igual a 1 indicará una diferencia pequeña.

Teorema 5-8: Sean dos muestras aleatorias independientes de tamaños m y n, respectivamente, que se obtienen de poblaciones normales con varianzas σ_1^2, σ_2^2, respectivamente. Entonces, si las varianzas de las muestras aleatorias están dadas por S_1^2, S_2^2, respectivamente, el estadístico

$$F = \frac{mS_1^2/(m-1)\sigma_1^2}{nS_2^2/(n-1)\sigma_2^2} = \frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2} \tag{23}$$

tiene distribución F con $m - 1$, $n - 1$ grados de libertad.

OTROS ESTADÍSTICOS

Se pueden definir muchos otros estadísticos, además de la media y la varianza o la desviación estándar. La mediana, la moda, los momentos, el sesgo y la curtosis son ejemplos. Sus definiciones son análogas a las dadas para las poblaciones en el capítulo 3. Con frecuencia se pueden encontrar las distribuciones muestrales para estos estadísticos, o al menos sus medias y desviaciones estándar (errores estándar). Algunos de éstos, junto con otros que ya dimos, aparecen en la tabla 5-1.

Tabla 5-1 Errores estándar para algunos estadísticos muestrales

Estadístico muestral	Error estándar	Observaciones
Medias	$\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$	Esto es verdadero para muestras grandes o pequeñas en donde la población es infinita o se hace muestreo con remplazo. La distribución muestral de medias es casi normal (normal asintóticamente) para $n \geqq 30$, incluso cuando la población no es normal. $\mu_{\bar{X}} = \mu$, la media de la población en todos los casos.
Proporciones	$\sigma_P = \sqrt{\dfrac{p(1-p)}{n}} = \sqrt{\dfrac{pq}{n}}$	Las observaciones hechas a las medias también aplican aquí. $\mu_p = p$ en todos los casos
Medianas	$\sigma_{\text{med}} = \sigma\sqrt{\dfrac{\pi}{2n}} = \dfrac{1.2533\sigma}{\sqrt{n}}$	Para $n \geqq 30$, la distribución muestral de las medianas es casi normal. Los resultados dados son válidos sólo si la población es normal o aproximadamente normal. $\mu_{\text{med}} = \mu$
Desviaciones estándar	(*1*) $\sigma_S = \dfrac{\sigma}{\sqrt{2n}}$ (*2*) $\sigma_S = \sqrt{\dfrac{\mu_4 - \sigma^4}{4n\sigma^2}}$	Para $n \geqq 100$, la distribución muestral de S es casi normal. σ_s está dada por (*1*) sólo si la población es normal (o aproximadamente normal). Si la población es no normal se puede usar (*2*). Observe que (*2*) se reduce a (*1*) cuando $\mu_4 = 3\sigma^4$, lo cual es verdadero para las poblaciones normales. Para $n \geqq 100$, $\mu_s = \sigma^2$ muy aproximada
Varianzas	(*1*) $\sigma_{S^2} = \sigma^2\sqrt{\dfrac{2}{n}}$ (*2*) $\sigma_{S^2} = \sqrt{\dfrac{\mu_4 - \sigma^4}{n}}$	Las observaciones hechas a la desviación estándar aplican también aquí. Observe que (*2*) se convierte en (*1*) cuando la población es normal. $\mu_{S^2} = (n-1)\sigma^2/n$ lo cual es casi σ^2 para n grande ($n \geqq 30$).

DISTRIBUCIONES DE FRECUENCIA

Si una muestra (o incluso una población) es grande, es difícil observar las diversas características o calcular estadísticas tales como la media o la desviación estándar. Por esta razón, es útil organizar o agrupar los *datos originales*. Como ilustración, supongamos que una muestra consta de la estatura de 100 estudiantes varones de la universidad XYZ. Agrupamos los datos en *clases* o *categorías* y determinamos el número de individuos que pertenecen a cada clase, llamada *frecuencia de clase*. La agrupación resultante, tabla 5-2, se denomina *distribución de frecuencia* o *tabla de frecuencia*.

La primera clase o categoría, por ejemplo, consta de estaturas que van desde 60 a 62 pulgadas, indicado por 60-62, que se denomina *intervalo de clase*. Dado que hay 5 estudiantes cuya estatura pertenece a esta clase, la correspondiente frecuencia de clase es 5. Puesto que una estatura registrada como 60 pulgadas en realidad se encuentra entre 59.5 y 60.5 pulgadas, mientras que otra registrada como 62 pulgadas en realidad se encuentra entre 61.5 y 62.5 pulgadas, podríamos también haber registrado el intervalo de clase como 59.5-62.5. El siguiente intervalo de clase podría ser, entonces, 62.5-65.5, etc. En el intervalo de clase 59.5-62.5 los números 59.5 y 62.5 se denominan *límite de clase*. El ancho del j-ésimo intervalo de clase, denotado por c_j, el cual es generalmente el mismo para todas las clases (en cuyo caso se denota por c), es la diferencia entre los límites de clase superior e inferior. En este caso $c = 62.5 - 59.5 = 3$.

El punto medio del intervalo de clase, que puede tomarse como representativo de la clase, se llama *marca de clase*. En la tabla anterior la marca de clase que corresponde al intervalo de clase 60-62 es 61.

Una gráfica para la distribución de frecuencia se puede suplir por un histograma, como se muestra en la parte sombreada de la figura 5-1, o por una *gráfica de polígono* (llamada *polígono de frecuencias*), que conecta los puntos medios de la parte superior del histograma. Es interesante ver que la forma de la gráfica parece indicar que la muestra se obtuvo de una población cuyas estaturas tienen distribución normal.

DISTRIBUCIONES DE FRECUENCIA RELATIVA

Si registramos en la tabla 5-2 la frecuencia relativa o de porcentaje en vez del número de estudiantes en cada clase, el resultado será una *distribución de frecuencia relativa* o de *porcentaje*. Por ejemplo, la distribución de frecuencia relativa o de porcentaje correspondiente a la clase 63-65 es 18/100, o 18%. El histograma correspondiente es entonces similar al de la figura 5-1, excepto porque el eje vertical es la frecuencia relativa en lugar de la frecuencia. La suma de las áreas rectangulares es entonces 1, o 100%.

Podemos considerar una distribución de frecuencia relativa como la distribución de probabilidad en la cual las probabilidades se remplazan por frecuencias relativas. Puesto que se puede pensar en las frecuencias relativas como *probabilidades empíricas* (*véase* página 5), las distribuciones de frecuencia relativa se conocen como *distribuciones de probabilidad empírica*.

Tabla 5-2
Estatura de 100 estudiantes varones de la Universidad XYZ

Estatura (pulgadas)	Número de Estudiantes
60–62	5
63–65	18
66–68	42
69–71	27
72–74	8
TOTAL	100

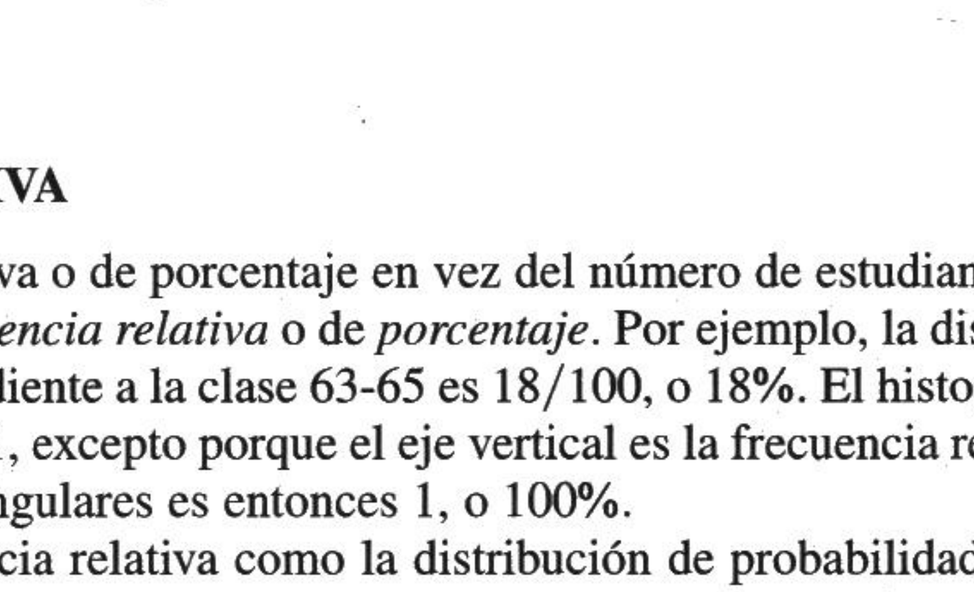


Fig. 5-1

CÁLCULO DE MEDIA, VARIANZA Y MOMENTOS PARA DATOS AGRUPADOS

Podemos representar una distribución de frecuencia como en la tabla 5-3, dando cada marca de clase con su correspondiente frecuencia de clase. La frecuencia total es n, es decir,

$$n = f_1 + f_2 + \cdots + f_k = \sum f$$

Tabla 5-3

Marca de clase	Frecuencia de clase
x_1	f_1
x_2	f_2
$\vdots$	$\vdots$
x_k	f_k
TOTAL	n

Ya que hay f_1 números iguales a x_1, f_2 números iguales a x_2, ..., f_k números iguales a x_k, la media está dada por

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_k x_k}{n} = \frac{\sum fx}{n} \tag{24}$$

De manera similar la varianza está dada por

$$s^2 = \frac{f_1(x_1 - \bar{x})^2 + f_2(x_2 - \bar{x})^2 + \cdots + f_k(x_k - \bar{x})^2}{n} = \frac{\sum f(x - \bar{x})^2}{n} \tag{25}$$

Observe la analogía de (*24*) y (*25*) con los resultados (*2*), página 82, y (*13*), página 84, si f_1/n corresponde a probabilidades empíricas.

En el caso donde todos los intervalos de clase tienen el mismo tamaño c, hay disponibles métodos cortos para calcular la media y la varianza. Éstos se llaman *métodos abreviados* y utilizan la transformación de la marca de clase x al entero correspondiente u dado por

$$x = a + cu \tag{26}$$

donde a es una marca de clase que se escoge arbitrariamente y corresponde a $u = 0$. Las fórmulas abreviadas para la media y la varianza están dadas por

$$\bar{x} = a + \frac{c}{n}\sum fu = a + c\bar{u} \tag{27}$$

$$s^2 = c^2\left[\frac{\sum fu^2}{n} - \left(\frac{\sum fu}{n}\right)^2\right] = c^2(\overline{u^2} - \bar{u}^2) \tag{28}$$

Para momentos superiores hay disponibles fórmulas similares. Los r-ésimos momentos alrededor de la media y el origen, respectivamente, están dados por

$$m_r = \frac{f_1(x_1 - \bar{x})^r + \cdots + f_k(x_k - \bar{x})^r}{n} = \frac{\sum f(x - \bar{x})^r}{n} \tag{29}$$

$$m_r' = \frac{f_1 x_1^r + \cdots + f_k x_k^r}{n} = \frac{\sum fx^r}{n} \tag{30}$$

Los dos tipos de momentos se relacionan por

$$\begin{aligned} m_1 &= 0 \\ m_2 &= m_2' - m_1'^2 \\ m_3 &= m_3' - 3m_1'm_2' + 2m_1'^3 \\ m_4 &= m_4' - 4m_1'm_3' + 6m_1'^2m_2' - 3m_1'^4 \end{aligned} \tag{31}$$

etc. Si escribimos

$$M_r = \frac{\sum f(u-\bar{u})^r}{n} \qquad M_r' = \frac{\sum fu^r}{n}$$

donde u está dada por (*26*), entonces las relaciones (*31*) también son válidas entre las M. Pero

$$m_r = \frac{\sum f(x-\bar{x})^r}{n} = \frac{\sum f[(a+cu)-(a+c\bar{u})]^r}{n} = \frac{\sum fc^r(u-\bar{u})^r}{n} = c^r M_r$$

así que obtenemos a partir de (*31*) las fórmulas abreviadas

$$\begin{aligned} m_1 &= 0 \\ m_2 &= c^2(M_2' - M_1'^2) \\ m_3 &= c^3(M_3' - 3M_1'M_2' + 2M_1'^3) \\ m_4 &= c^4(M_4' - 4M_1'M_3' + 6M_1'^2M_2' - 3M_1'^4) \end{aligned} \tag{32}$$

etc. La segunda ecuación (*32*) es, por supuesto, la misma (*28*).

De manera similar para muestras agrupadas se pueden encontrar otras estadísticas, como el sesgo y la curtosis.

Problemas resueltos

DISTRIBUCIÓN MUESTRAL DE MEDIAS

5.1. Una población consta de cinco números 2, 3, 6, 8, 11. Considere todas las muestras posibles de tamaño dos que pueden obtenerse con remplazo de esta población. Encuentre (*a*) la media de la población, (*b*) la desviación estándar de la población, (*c*) la media de la distribución muestral de medias, (*d*) la desviación estándar de la distribución muestral de medias, es decir, el error estándar de las medias.

(*a*)
$$\mu = \frac{2+3+6+8+11}{5} = \frac{30}{5} = 6.0$$

(*b*)
$$\sigma^2 = \frac{(2-6)^2+(3-6)^2+(6-6)^2+(8-6)^2+(11-6)^2}{5} = \frac{16+9+0+4+25}{5} = 10.8$$

y $\sigma = 3.29$.

(*c*) Hay $5(5) = 25$ muestras de tamaño dos que se pueden obtener con remplazo (ya que cualquiera de los cinco números en el primer retiro puede asociarse con cualquiera de los cinco números en el segundo retiro). Éstos son

$$\begin{array}{lllll} (2,2) & (2,3) & (2,6) & (2,8) & (2,11) \\ (3,2) & (3,3) & (3,6) & (3,8) & (3,11) \\ (6,2) & (6,3) & (6,6) & (6,8) & (6,11) \\ (8,2) & (8,3) & (8,6) & (8,8) & (8,11) \\ (11,2) & (11,3) & (11,6) & (11,8) & (11,11) \end{array}$$

Las medias muestrales correspondientes son

$$(1) \qquad \begin{array}{lllll} 2.0 & 2.5 & 4.0 & 5.0 & 6.5 \\ 2.5 & 3.0 & 4.5 & 5.5 & 7.0 \\ 4.0 & 4.5 & 6.0 & 7.0 & 8.5 \\ 5.0 & 5.5 & 7.0 & 8.0 & 9.5 \\ 6.5 & 7.0 & 8.5 & 9.5 & 11.0 \end{array}$$

y la media de la distribución muestral de medias es

$$\mu_{\bar{X}} = \frac{\text{suma de todas las medias muestrales en } (1) \text{ arriba}}{25} = \frac{150}{25} = 6.0$$

ilustrando el hecho de que $\mu_{\bar{X}} = \mu$. Para una demostración general de esto, *véase* el problema 5.6.

(*d*) La varianza $\sigma_{\bar{X}}^2$ de la distribución muestral de medias se obtiene restando la media 6 de cada número en (*1*), elevando al cuadrado el resultado, sumando los 25 números obtenidos y dividiendo por 25. El resultado final es

$$\sigma_{\bar{X}}^2 = \frac{135}{25} = 5.40 \text{ de manera que } \sigma_{\bar{X}} = \sqrt{5.40} = 2.32$$

Esto ilustra el hecho de que para poblaciones finitas que involucran el muestreo con remplazo (o poblaciones infinitas), $\sigma_{\bar{X}}^2 = \sigma^2/n$ dado que el lado derecho es $10.8/2 = 5.40$, lo que concuerda con el valor de arriba. Para una demostración general de esto, *véase* el problema 5.7.

5.2. Resuelva el problema 5.1 para el caso de muestreo sin remplazo.

Tal como en (*a*) y (*b*) del problema 5.1, $\mu = 6$ y $\sigma^2 = 10.8$, $\sigma = 3.29$.

(*c*) Hay ${}_5C_2 = 10$ muestras de tamaño dos que pueden obtenerse sin remplazo (esto significa que sacamos un número y luego otro diferente del primero) de la población, a saber,

$$(2,3), \quad (2,6), \quad (2,8), \quad (2,11), \quad (3,6), \quad (3,8), \quad (3,11), \quad (6,8), \quad (6,11), \quad (8,11)$$

La selección (2, 3), por ejemplo, se considera la misma que (3, 2)

Las correspondientes medias muestrales son

$$2.5, \quad 4.0, \quad 5.0, \quad 6.5, \quad 4.5, \quad 5.5, \quad 7.0, \quad 7.0, \quad 8.5, \quad 9.5$$

y la media de la distribución muestral de medias es

$$\mu_{\bar{X}} = \frac{2.5 + 4.0 + 5.0 + 6.5 + 4.5 + 5.5 + 7.0 + 7.0 + 8.5 + 9.5}{10} = 6.0$$

ilustrando el hecho de que $\mu_{\bar{X}} = \mu$.

(*d*) La varianza de la distribución muestral de medias es

$$\sigma_{\bar{X}}^2 = \frac{(2.5 - 6.0)^2 + (4.0 - 6.0)^2 + (5.0 - 6.0)^2 + \cdots + (9.5 - 6.0)^2}{10} = 4.05$$

y $\sigma_{\bar{X}} = 2.01$.

Esto ilustra $\sigma_{\bar{X}}^2 = \dfrac{\sigma^2}{n}\left(\dfrac{N-n}{N-1}\right)$, ya que el lado derecho es igual a $\dfrac{10.8}{2}\left(\dfrac{5-2}{5-1}\right) = 4.05$, tal como se obtuvo atrás. Para una demostración general de este resultado, *véase* el problema 5.47.

5.3. Suponga que las estaturas de 3000 estudiantes varones en una universidad tiene distribución normal con media 68.0 pulgadas y desviación estándar 3.0 pulgadas. Si se obtienen 80 muestras de 25 estudiantes cada una, ¿cuál será la media y la desviación estándar de la distribución muestral de medias si las muestras se tomaron (*a*) con remplazo, (*b*) sin remplazo?

El número de muestras de tamaño 25, que teóricamente se pueden obtener de un grupo de 3000 estudiantes con y sin remplazo son $(3000)^{25}$ y ${}_{3000}C_{25}$, los cuales son mayores que 80. Así que no obtenemos una verdadera distribución muestral de medias sino sólo una *distribución muestral experimental*. En todo caso, dado que el número de muestras es grande, habrá una coincidencia cercana entre las dos distribuciones muestrales. Así, la media y la desviación estándar de las 80 muestras estarán cerca de la distribución teórica. Por consiguiente, tenemos

(*a*) $\mu_{\bar{X}} = \mu =$ pulgadas y $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{3}{\sqrt{25}} = 0.6$ pulgadas

(*b*) $\mu_{\bar{X}} = \mu = 68.0$ pulgadas y $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}} = \dfrac{3}{\sqrt{25}}\sqrt{\dfrac{3000-25}{3000-1}}$

lo cual es tan sólo ligeramente menor que 0.6 pulgadas y puede considerarse para todo propósito práctico como igual, tal como en el muestreo con remplazo.

Es así como podemos esperar que la distribución muestral de medias experimental tenga distribución aproximadamente normal con media 68.0 pulgadas y desviación estándar 0.6 pulgadas.

5.4. ¿En cuántas muestras del problema 5.3 esperaría encontrar la media (*a*) entre 66.8 y 68.3 pulgadas, (*b*) menor que 66.4 pulgadas?

La media $\bar{X}$ de una muestra en unidades estándar está dada por $Z = \dfrac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \dfrac{\bar{X} - 68.0}{0.6}$.

(*a*) 66.8 en unidades estándar = (66.8 − 68.0) / 0.6 = −2.0
66.3 en unidades estándar = (66.3 − 68.0) / 0.6 = 0.5

Proporción de muestras con medias entre 66.8 y 68.3 pulgadas
= (área bajo la curva normal entre $z = -2.0$ y $z = 0.5$)
= (área entre $z = -2$ y $z = 0$)
+ (área entre $z = 0$ y $z = 0.5$)
= 0.4772 + 0.1915 = 0.6687

Entonces el número esperado de muestras = (80)(0.6687) o 53 (figura 5-2).

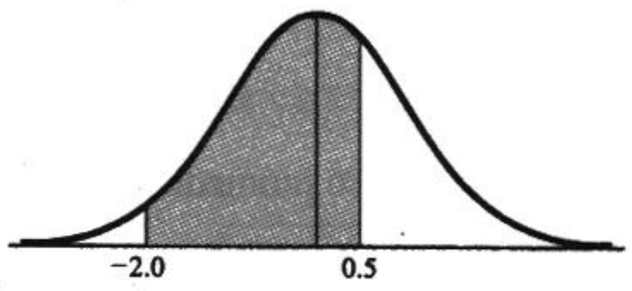


Fig. 5-2

(*b*) 66.4 en unidades estándar = (66.4 − 68.0)/0.6 = −2.67

Proporción de muestras con medias menores que 66.4 pulgadas

= (área bajo la curva normal a la izquierda de $z = -2.67$)

= (área a la izquierda de $z = 0$)

− (área entre $z = -2.67$ y $z = 0$)

= 0.5 − 0.4962 = 0.0038

Entonces el número esperado de muestras = (80)(0.0038) = 0.304 o cero (figura 5-3).

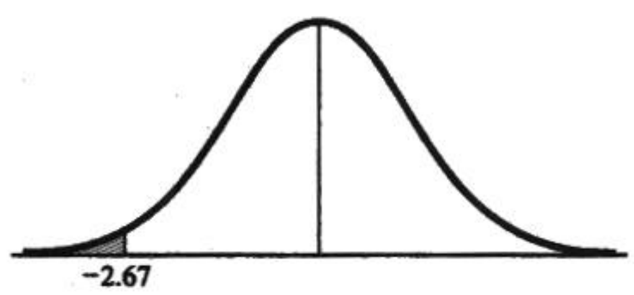


Fig. 5-3

5.5. Quinientas balineras tienen un peso promedio de 5.02 gramos y una desviación estándar de 0.30 gramos. Encuentre la probabilidad de que una muestra aleatoria de 100 balineras escogidas de este grupo tengan un peso combinado (*a*) entre 496 y 500 gramos, (*b*) de más de 510 gramos.

Para la distribución muestral de medias, $\mu_{\bar{X}} = \mu = 5.02$ gramos.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{0.30}{\sqrt{100}}\sqrt{\frac{500-100}{500-1}} = 0.027$$

(*a*) El peso combinado estará entre 496 y 500 gramos si el peso promedio de 100 balineras se encuentra entre 4.96 y 5.00 gramos (figura 5-4).

$$\textbf{4.96 en unidades estándar} = \frac{4.96 - 5.02}{0.027} = -2.22$$

$$\textbf{5.00 en unidades estándar} = \frac{5.00 - 5.02}{0.027} = -0.74$$

Probabilidad requerida

= (área entre $z = -2.22$ y $z = -0.74$)

= (área entre $z = -2.22$ y $z = 0$)

− (área entre $z = -0.74$ y $z = 0$)

= 0.4868 − 0.2704 = 0.2164

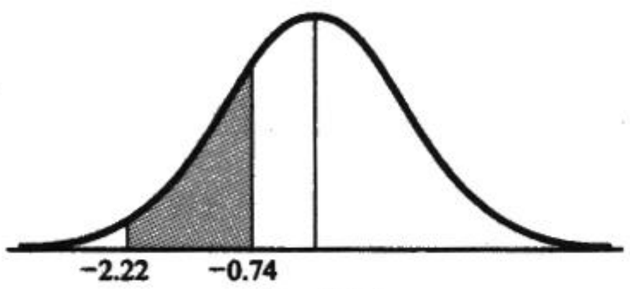


Fig. 5-4

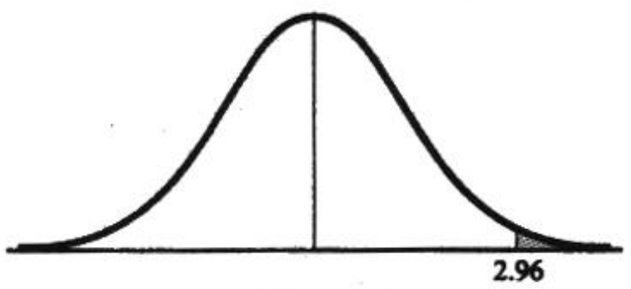


Fig. 5-5

(*b*) El peso combinado excederá 510 gramos si el peso promedio de las 100 balineras excede 5.10 gramos (figura 5-5).

$$\textbf{5.10 en unidades estándar} = \frac{5.10 - 5.02}{0.027} = 2.96$$

Probabilidad requerida
= (área a la derecha de $z = 2.96$)
= (área a la derecha de $z = 0$)
− (área entre $z = 0$ y $z = 2.96$)
$= 0.5 - 0.4985 = 0.0015$

Por consiguiente hay sólo 3 posibilidades en 2000 de obtener una muestra de 100 balineras con un peso combinado que exceda 510 gramos.

5.6. Demuestre el teorema 5-1, página 168.

Dado que $X_1, X_2, \ldots, X_n$ son variables aleatorias con la misma distribución de la población, la cual tiene media μ, tenemos

$$E(X_k) = \mu \qquad k = 1, 2, \ldots, n$$

Entonces, dado que la media muestral se define como

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$$

tenemos, como se requiere

$$E(\bar{X}) = \frac{1}{n}[E(X_1) + \cdots + E(X_n)] = \frac{1}{n}(n_\mu) = \mu$$

5.7. Demuestre el teorema 5-2, página 168.

Tenemos

$$\bar{X} = \frac{X_1}{n} + \frac{X_2}{n} + \cdots + \frac{X_n}{n}$$

Entonces dado que $X_1, \ldots, X_n$ son variables independientes y tienen varianza σ^2, tenemos de los teoremas 3-5 y 3-7:

$$\text{Var}(\bar{X}) = \frac{1}{n^2}\text{Var}(X_1) + \cdots + \frac{1}{n^2}\text{Var}(X_n) = n\left(\frac{1}{n^2}\sigma^2\right) = \frac{\sigma^2}{n}$$

DISTRIBUCIÓN MUESTRAL DE PROPORCIONES

5.8. Encuentre la probabilidad de que en 120 lanzamientos de una moneda balanceada (*a*) entre 40% y 60% sean caras, (*b*) $\frac{5}{8}$ o más sean caras.

Consideramos los 120 lanzamientos de la moneda como una muestra de la población infinita de sus posibles lanzamientos. En esta población la probabilidad de caras es $p = \frac{1}{2}$ y la probabilidad de sellos es $q = 1 - p = \frac{1}{2}$.

(*a*) Requerimos la probabilidad de que el número de caras en 120 lanzamientos esté entre 40% de 120, o 48, y 60% de 120, o 72. Procedemos como en el capítulo 4, usando la aproximación normal a la distribución binomial. Dado que el número de caras es una variable discreta, queremos saber la probabilidad de que el número de caras esté entre 47.5 y 72.5 (*véase* fig. 5-6.)

$$\mu = \text{número esperado de caras} = np = 120\left(\frac{1}{2}\right) = 60$$

y

$$\sigma = \sqrt{npq} = \sqrt{(120)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)} = 5.48$$

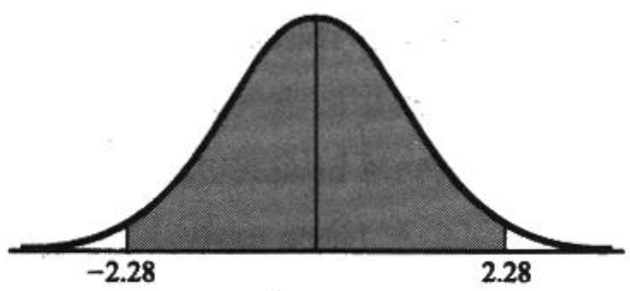


Fig. 5-6

$$47.5 \text{ en unidades estándar} = \frac{47.5 - 60}{5.48} = -2.28$$

$$72.5 \text{ en unidades estándar} = \frac{72.5 - 60}{5.48} = 2.28$$

Probabilidad requerida

= (área bajo la curva normal entre $z = -2.28$ y $z = 2.28$)

= 2(área entre $z = 0$ y $z = 2.28$)

$= 2(0.4887) = 0.9774$

Otro método.

$$\mu_P = p = \frac{1}{2} = 0.50 \qquad \sigma_P = \sqrt{\frac{pq}{n}} = \sqrt{\frac{\frac{1}{2}\left(\frac{1}{2}\right)}{120}} = 0.0456$$

$$40\% \text{ en unidades estándar} = \frac{0.40 - 0.50}{0.0456} = -2.19$$

$$60\% \text{ en unidades estándar} = \frac{0.60 - 0.50}{0.0456} = 2.19$$

Por consiguiente, la probabilidad requerida es el área bajo la curva normal entre $z = -2.19$ y $z = 2.19$, es decir, $2(0.4857) = 0.9714$.

Aunque este resultado tiene dos cifras significativas, no coincide exactamente ya que no usamos el hecho de que la proporción es en realidad una variable discreta. Teniendo en cuenta esto, restamos $\frac{1}{2n} = \frac{1}{2(120)}$ de 0.4 y sumamos $\frac{1}{2n} = \frac{1}{2(120)}$ a 0.60. Por consiguiente, las proporciones requeridas en unidades estándar son, dado que $1/240 = 0.00417$,

$$\frac{0.40 - 0.00417 - 0.50}{0.0456} = -2.28 \qquad \text{y} \qquad \frac{0.60 + 0.00417 - 0.50}{0.0456} = 2.28$$

de manera que concuerda con lo obtenido en el primer método.

Observe que $(0.40 - 0.00417)$ y $(0.60 + 0.00417)$ corresponden a las proporciones $47.5/120$ y $72.5/120$ del método anterior.

(*b*) Usando el segundo método de (*a*), encontramos que $\frac{5}{8} = 0.6250$,

$$(0.6250 - 0.00417) \text{ en unidades estándar} = \frac{0.6250 - 0.00417 - 0.50}{0.0456} = 2.65$$

Probabilidad requerida = (área bajo la curva normal a la derecha de $z = 2.65$)

= (área a la derecha de $z = 0$)

− (área entre $z = 0$ y $z = 2.65$)

$= 0.5 - 0.4960 = 0.0040$

5.9. Cada persona de un grupo de 500 lanza una moneda balanceada 120 veces. ¿Cuántas personas esperan reportar que (*a*) entre 40% y 60% de sus lanzamientos fueron caras, (*b*) $\frac{5}{8}$ o más de sus lanzamientos fueron caras?

Este problema está muy relacionado con el problema 5.8. Aquí consideramos 500 muestras, cada una de tamaño 120, de la población infinita de los posibles lanzamientos de una moneda.

(*a*) La parte (*a*) del problema 5.8 establece que de las muestras posibles, cada una de 120 lanzamientos de una moneda, podemos esperar encontrar 97.74% con un porcentaje de caras entre 40% y 60%. En 500 muestras podemos esperar encontrar alrededor de 97.74% de 500, o 489, muestras con esta propiedad. Se deduce que cerca de 489 personas podrán esperar reportar que su experimento resultó entre 40% y 60% de caras.

Es interesante observar que $500 - 489 = 11$ personas pueden esperar que el porcentaje de caras no esté entre 40% y 60%. Estas personas pueden concluir, de manera razonable, que sus monedas estaban cargadas, incluso si éstas fueran balanceadas. Este tipo de error es un *riesgo* que siempre está presente mientras tratemos con la probabilidad.

(*b*) Razonando como en (*a*) concluimos que cerca de $(500)(0.0040) = 2$ personas reportarán que $\frac{5}{8}$ o más de sus lanzamientos dieron como resultado caras.

5.10. Se ha encontrado que 2% de las herramientas producidas por cierta máquina son defectuosas. ¿Cuál es la probabilidad de que de un cargamento de 400 de esas herramientas sean defectuosas (*a*) 3% o más, (*b*) 2% o menos?

$$\mu_P = p = 0.02 \quad \text{y} \quad \sigma_P = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.02(0.98)}{400}} = \frac{0.14}{20} = 0.007$$

(*a*) Usando la corrección para variables discretas, $1/2n = 1/800 = 0.00125$, tenemos,

$$(0.03 - 0.00125) \text{ en unidades estándar} = \frac{0.03 - 0.00125 - 0.02}{0.007} = 1.25$$

Probabilidad requerida = (área bajo la curva normal a la derecha de $z = 1.25$) = 0.1056

Si no usamos la corrección hubiéramos obtenido 0.0764.

Otro método.

(3% de 400) = 12 herramientas defectuosas. Sobre una base continua, 12 o más herramientas significan 11.5 o más.

$$\mu = (2\% \text{ de } 400) = 8 \quad ; \quad \sigma = \sqrt{npq} = \sqrt{(400)(0.02)(0.98)} = 2.8$$

Entonces, 11.5 en unidades estándar = $(11.5 - 8)/2.8 = 1.25$, y tal como en el método anterior, la probabilidad requerida es 0.1056.

(*b*)
$$(0.02 + 0.00125) \text{ en unidades estándar} = \frac{0.02 + 0.00125 - 0.02}{0.007} = 0.18$$

Probabilidad requerida = (área bajo la curva normal a la izquierda de $z = 0.18$)
$= 0.5000 + 0.0714 = 0.5714$

Si no usamos la corrección, hubiéramos obtenido 0.5000. También se puede usar el segundo método de la parte (*a*).

5.11. Los resultados de una elección demostraron que cierto candidato recibió el 46% de los votos. Determine la probabilidad de que una encuesta de (*a*) 200, (*b*) 1000 personas de la población votante seleccionadas al azar de la población votante haya mostrado mayoría de votos a favor del candidato.

(*a*)
$$\mu_P = p = 0.46 \quad \text{y} \quad \sigma_P = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.46(0.54)}{200}} = 0.0352$$

Dado que $1/2n = 1/400 = 0.0025$, una mayoría se indica en la muestra si la proporción a favor del candidato es $0.50 + 0.0025 = 0.5025$ o más (también se puede obtener esta proporción si nos damos cuenta que 101 o más indica mayoría, pero ésta como variable continua es 100.5; y así, la proporción es $100.5/200 = 0.5025$).

Entonces, 0.5025 en unidades estándar = (0.5025 − 0.46)/0.0352 = 1.21 y

Probabilidad requerida = (área bajo la curva normal a la derecha de $z = 1.21$)

= 0.5000 − 0.3869 = 0.1131

(*b*) $\mu_P = p = 0.46, \sigma_P = \sqrt{pq/n} = \sqrt{0.46(0.54)1000} = 0.0158$, y

$$0.5025 \text{ en unidades estándar } = \frac{0.5025 - 0.46}{0.0158} = 2.69$$

Probabilidad requerida = (área bajo la curva normal a la derecha de $z = 2.69$)

= 0.5000 − 0.4964 = 0.0036

DISTRIBUCIÓN MUESTRAL DE DIFERENCIAS Y SUMAS

5.12. Sea U_1 una variable que significa cualquiera de los elementos de la población 3, 7, 8 y U_2 una variable que significa cualquiera de los elementos 2, 4. Calcule (*a*) μ_{U_1}, (*b*) μ_{U_2}, (*c*) $\mu_{U_1-U_2}$, (*d*) σ_{U_1}, (*e*) σ_{U_2}, (*f*) $\sigma_{U_1-U_2}$.

(*a*) $$\mu_{U_1} = \text{ media de la población } U_1 = \frac{1}{3}(3+7+8) = 6$$

(*b*) $$\mu_{U_2} = \text{ media de la población } U_2 = \frac{1}{2}(2+4) = 3$$

(*c*) La población que consta de las diferencias de cualquier miembro de U_1 y cualquier miembro de U_2 es

$$\begin{matrix} 3-2 & 7-2 & 8-2 \\ 3-4 & 7-4 & 8-4 \end{matrix} \quad \text{o} \quad \begin{matrix} 1 & 5 & 6 \\ -1 & 3 & 4 \end{matrix}$$

Entonces $$\mu_{U_1-U_2} = \text{media de } (U_1 - U_2) = \frac{1+5+6+(-1)+3+4}{6} = 3$$

lo cual ilustra el resultado general $\mu_{U_1-U_2} = \mu_{U_1-U_2}$, como se vio en (*a*) y (*b*).

(*d*) $$\sigma^2_{U_1} = \text{ varianza de la población } U_1 = \frac{(3-6)^2 + (7-6)^2 + (8-6)^2}{3} = \frac{14}{3}$$

o $\sigma_{U_1} = \sqrt{\frac{14}{3}}$.

(*e*) $$\sigma^2_{U_2} = \text{ varianza de la población } U_2 = \frac{(2-3)^2 + (4-3)^2}{2} = 1$$

o $\sigma_{U_2} = 1$.

(*f*) $$\sigma^2_{U_1-U_2} = \text{ varianza de la población } (U_1 - U_2)$$

$$= \frac{(1-3)^2 + (5-3)^2 + (6-3)^2 + (-1-3)^2 + (3-3)^2 + (4-3)^2}{6} = \frac{17}{3}$$

o $\sigma_{U_1-U_2} = \sqrt{\frac{17}{3}}$.

Esto ilustra el resultado general para las muestras independientes, $\sigma_{U_1-U_2} = \sqrt{\sigma^2_{U_1} + \sigma^2_{U_2}}$, como se vio en (*d*) y (*e*). La demostración para el resultado general se deduce del teorema 3-7, página 85.

5.13. Las bombillas eléctricas del fabricante *A* tienen una vida media de 1400 horas con una desviación estándar de 200 horas, mientras que las hechas por el fabricante *B* tienen una vida media de 1200 horas con desviación estándar de 100 horas. Si se toman muestras al azar de 125 bombillas de cada marca que se está probando, ¿cuál es la probabilidad de que la marca *A* tenga un tiempo de vida media de al menos (*a*) 160 horas, (*b*) 250 horas más que las bombillas de la marca *B*?

Denoten $\bar{X}_A$ y $\bar{X}_B$ los tiempos de vida media de las muestras A y B, respectivamente. Entonces

$$\mu_{\bar{X}_A-\bar{X}_B} = \mu_{\bar{X}_A} - \mu_{\bar{X}_B} = 1400 - 1200 = 200 \text{ horas}$$

y

$$\sigma_{\bar{X}_A-\bar{X}_B} = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} = \sqrt{\frac{(100)^2}{125} + \frac{(200)^2}{125}} = 20 \text{ horas}$$

La variable estandarizada para la diferencia de medias

$$Z = \frac{(\bar{X}_A - \bar{X}_B) - (\mu_{\bar{X}_A-\bar{X}_B})}{\sigma_{\bar{X}_A-\bar{X}_B}} = \frac{(\bar{X}_A - \bar{X}_B) - 200}{20}$$

tiene casi una distribución normal.

(a) La diferencia de 160 horas en unidades estándar $= (160 - 200)/20 = -2$

Probabilidad requerida = (área bajo la curva normal a la derecha de $z = -2$)

$= 0.5000 + 0.4772 = 0.9772$

(b) La diferencia de 250 horas en unidades estándar $= (250 - 200)/20 = 2.50$

Probabilidad requerida = (área bajo la curva normal a la derecha de $z = 2.50$)

$= 0.5000 - 0.4938 = 0.0062$

5.14. Las balineras de cierta marca pesan 0.50 gramos con desviación estándar de 0.02 gramos. ¿Cuál es la probabilidad de que dos lotes de 1000 balineras cada uno difieran en peso por más de 2 gramos?

Sean $\bar{X}_1$ y $\bar{X}_2$ los pesos promedio de las balineras de dos lotes. Entonces

$$\mu_{\bar{X}_1-\bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = 0.50 - 0.50 = 0$$

$$\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(0.02)^2}{1000} + \frac{(0.02)^2}{1000}} = 0.000895$$

La variable estandarizada para la diferencia de medias es $Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - 0}{0.000895}$ y casi tiene distribución normal.

La diferencia de 2 gramos en los lotes equivale a una diferencia de $2/1000 = 0.002$ gramos en las medias. Esto puede ocurrir si $\bar{X}_1 - \bar{X}_2 \geqq 0.002$ o $\bar{X}_1 - \bar{X}_2 \leqq -0.002$, es decir;

$$Z \geqq \frac{0.002 - 0}{0.000895} = 2.23 \quad \text{o} \quad Z \leqq \frac{-0.002 - 0}{= 0.000895} = -2.23$$

Entonces

$$P(Z \geqq 2.23 \text{ or } Z \leqq -2.23) = P(Z \geqq 2.23) + P(Z \leqq -2.23) = 2(0.5000 - 0.4871) = 0.0258$$

5.15. A y B juegan a cara y sello, cada uno lanzando 50 monedas. A gana el juego si lanza 5 caras o más que B, de lo contrario gana B. Determine las posibilidades en contra de que A gane cualquier juego en particular.

Sean P_A y P_B las proporciones de caras obtenidas por A y B. Si suponemos que las monedas son balanceadas, la probabilidad p de caras es $\frac{1}{2}$. Entonces

$$\mu_{P_A-P_B} = \mu_{P_A} - \mu_{P_B} = 0$$

y

$$\sigma_{P_A-P_B} = \sqrt{\sigma_{P_A}^2 + \sigma_{P_B}^2} = \sqrt{\frac{pq}{n_A} + \frac{pq}{n_B}} = \sqrt{\frac{2(\frac{1}{2})(\frac{1}{2})}{50}} = 0.10$$

La variable estandarizada para la diferencia de proporciones es $Z = (P_A - P_B - 0)/0.10$.

Sobre la base de una variable continua, 5 o más caras significan 4.5 o más caras, de manera que la diferencia de proporciones debe ser $4.5/50 = 0.009$ o más, es decir, Z es mayor o igual que $(0.09 - 0)/0.10 = 0.9$ (o $Z \geqq 0.9$). La probabilidad de esto es el área bajo la curva normal a la derecha de $Z = 0.9$, lo cual es $0.5000 - 0.3159 = 0.1841$.

Por consiguiente, las posibilidades en contra de que A gane son $(1 - 0.1841) : 0.1841 = 0.8159 : 0.1841$ o 4.43 a 1.

5.16. Se miden dos distancias: 27.3 metros y 15.6 metros, con desviaciones estándar (errores estándar) de 0.16 metros y 0.08 metros, respectivamente. Determine la media y la desviación estándar de (*a*) la suma, (*b*) la diferencia de las distancias.

Si las distancias se denominan D_1 y D_2, entonces

(*a*)
$$\mu_{D_1+D_2} = \mu_{D_1} + \mu_{D_2} = 27.3 + 15.6 = 42.9 \text{ metros}$$
$$\sigma_{D_1+D_2} = \sqrt{\sigma_{D_1}^2 + \sigma_{D_2}^2} = \sqrt{(0.16)^2 + (0.08)^2} = 0.18 \text{ metros}$$

(*b*)
$$\mu_{D_1-D_2} = \mu_{D_1} + \mu_{D_2} = 27.3 + 15.6 = 11.7 \text{ metros}$$
$$\sigma_{D_1-D_2} = \sqrt{\sigma_{D_1}^2 + \sigma_{D_2}^2} = \sqrt{(0.16)^2 + (0.08)^2} = 0.18 \text{ metros}$$

5.17. Cierto tipo de bombilla eléctrica tiene una vida media de 1500 horas y una desviación estándar de 150 horas. Se conectan tres bombillas de manera que cuando una se funda, otra se encienda. Suponiendo que las vidas medias tienen distribución normal, ¿cuál es la probabilidad de que se tenga luz (*a*) al menos durante 5000 horas, (*b*) máximo durante 4200 horas?

Denominemos los tiempos de vida media como L_1, L_2 y L_3. Entonces

$$\mu_{L_1+L_2+L_3} = \mu_{L_1} + \mu_{L_2} + \mu_{L_3} = 1500 + 1500 + 1500 = 4500 \text{ horas}$$
$$\sigma_{L_1+L_2+L_3} = \sqrt{\sigma_{L_1}^2 + \sigma_{L_2}^2 + \sigma_{L_3}^2} = \sqrt{3(150)^2} = 260 \text{ horas}$$

(*a*) 5000 horas en unidades estándar $= (5000 - 4500)/260 = 1.92$

Probabilidad requerida = (área bajo la curva normal a la derecha de $z = 1.92$)
$= 0.5000 - 0.4726 = 0.0274$

(*b*) 4200 horas en unidades estándar $= (4200 - 4500)/260 = -1.15$

Probabilidad requerida = (área bajo la curva normal a la izquierda de $z = -1.15$)
$= 0.5000 - 0.3749 = 0.1251$

DISTRIBUCIÓN MUESTRAL DE VARIANZAS

5.18. Con relación al problema 5.1, encuentre (*a*) la media de la distribución muestral de varianzas, (*b*) la desviación estándar de la distribución muestral de varianzas, es decir, el error estándar de las varianzas.

(*a*) Las varianzas muestrales correspondientes a cada una de las 25 muestras del problema 5.1 son

0	0.25	4.00	9.00	20.25
0.25	0	2.25	6.25	16.00
4.00	2.25	0	1.00	6.25
9.00	6.25	1.00	0	2.25
20.25	16.00	6.25	2.25	0

La media de la distribución muestral de varianzas es

$$\mu_{S^2} = \frac{\text{suma de todas las varianzas de la tabla anterior}}{25} = \frac{135}{25} = 5.40$$

Esto ilustra el hecho de que $\mu_{S2} = (n-1)(\sigma^2)/n$ ya que $n = 2$ y $\sigma^2 = 10.8$ [*véase* el problema 5.1(*b*)], el lado derecho es $(\frac{1}{2})(10.8) = 5.4$.

El resultado indica por qué la varianza corregida para muestras se define con frecuencia como $\hat{S}^2 = \dfrac{n}{n-1} S^2$, ya que $\mu_{\hat{S}^2} = \sigma^2$ (*véanse también* las observaciones de la página 170).

(*b*) La varianza de la distribución muestral de varianzas $\sigma^2_{S^2}$ se obtiene restando la media 5.40 de cada uno de los 25 números de la tabla anterior, elevando estos números al cuadrado, sumándolos y luego dividiendo el resultado por 25. Por consiguiente, $\sigma^2_{S^2} = 575.75/25 = 23.03$ o $\sigma_{S^2} = 4.80$.

5.19. Desarrolle el problema anterior si se toman muestras sin remplazo.

(*a*) Hay 10 muestras cuyas varianzas están dadas por los números ubicados arriba (o abajo) de la diagonal de ceros en la tabla del problema 5.18(*a*). Entonces

$$\mu_{S^2} = \frac{0.25 + 4.00 + 9.00 + 20.25 + 2.25 + 6.25 + 16.00 + 1.00 + 6.25 + 2.25}{10} = 6.75$$

Éste es un caso especial del resultado general $\mu_{S^2} = \left(\dfrac{N}{N-1}\right)\left(\dfrac{n-1}{n}\right)\sigma^2$ [ecuación (*19*), página 170)]

como se comprueba colocando $N = 5$ y $\sigma^2 = 10.8$ en el lado derecho para obtener $\mu_{S^2} = (\frac{5}{4})(\frac{1}{2})(10.8) = 6.75$.

(*b*) Restando 6.75 a cada uno de los 10 números ubicados por encima de la diagonal de ceros en la tabla del problema 5.18(*a*), elevándolos al cuadrado, sumándolos, y dividiéndolos por 10, encontramos $\sigma^2_{S^2} = 39.675$ o $\sigma_{S^2} = 6.30$.

5.20. Demuestre que

$$E(S^2) = \frac{n-1}{n}\sigma^2$$

donde S^2 es la varianza muestral de una muestra aleatoria de tamaño n, como se definió en la página 170, y σ^2 es la varianza de la población.

Método 1.

Tenemos

$$X_1 - \bar{X} = X_1 - \frac{1}{n}(X_1 + \cdots + X_n) = \frac{1}{n}[(n-1)X_1 - X_2 - \cdots - X_n]$$
$$= \frac{1}{n}[(n-1)(X_1 - \mu) - (X_2 - \mu) - \cdots - (X_n - \mu)]$$

Entonces

$$(X_1 - \bar{X})^2 = \frac{1}{n^2}[(n-1)^2(X_1 - \mu)^2 + (X_2 - \mu)^2 + \cdots + (X_n - \mu)^2 + \text{términos de producto cruzado}]$$

Dado que los X son independientes, la esperanza de cada producto cruzado es cero, y tenemos

$$E[(X_1 - \bar{X})^2] = \frac{1}{n^2}\{(n-1)^2 E[(X_1 - \mu)^2] + E[(X_2 - \mu)^2] + \cdots + E[(X_n - \mu)^2]\}$$
$$= \frac{1}{n^2}\{(n-1)^2\sigma^2 + \sigma^2 + \cdots + \sigma^2\}$$
$$= \frac{1}{n^2}\{(n-1)^2\sigma^2 + (n-1)\sigma^2\} = \frac{n-1}{n}\sigma^2$$

De manera similar, $E[(X_k - \bar{X})^2] = (n-1)\sigma^2/n$ para $k = 2, \ldots, n$. Por consiguiente,

$$E(S^2) = \frac{1}{n} E[(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2]$$
$$= \frac{1}{n}\left[\frac{n-1}{n}\sigma^2 + \cdots + \frac{n-1}{n}\sigma^2\right] = \frac{n-1}{n}\sigma^2$$

Método 2.

Tenemos $X_j - \bar{X} = (X_j - \mu) - (\bar{X} - \mu)$. Entonces

$$(X_j - \bar{X})^2 = (X_j - \mu)^2 - 2(X_j - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2$$

y

$$(1) \qquad \sum (X_j - \bar{X})^2 = \sum (X_j - \mu)^2 - 2(\bar{X} - \mu)\sum (X_j - \mu) + \sum (\bar{X} - \mu)^2$$

donde la suma es desde $j = 1$ hasta n. Esto se puede escribir

$$(2) \qquad \sum (X_j - \bar{X})^2 = \sum (X_j - \mu)^2 - 2n(\bar{X} - \mu)^2 + n(\bar{X} - \mu)^2$$
$$= \sum (X_j - \mu)^2 - n(\bar{X} - \mu)^2$$

ya que $\sum (X_j - \mu) = \sum X_j - n_\mu = n(\bar{X} - \mu)$. Tomando la esperanza de ambos lados de (2) y usando el problema 5.7, encontramos

$$E[\sum (X_j - \bar{X})^2] = E[\sum (X_j - \mu)^2] - nE[(\bar{X} - \mu)^2]$$
$$= n\sigma^2 - n\left(\frac{\sigma^2}{n}\right) = (n-1)\sigma^2$$

de donde

$$E(S^2) = \frac{n-1}{n}\sigma^2$$

5.21. Demuestre el teorema 5-4, página 168.

Si X_j, $j = 1, 2, \ldots, n$, tiene distribución normal con media μ y varianza σ^2, entonces la función característica es (*véase* tabla 4-2, página 120)

$$\phi_j(\omega) = e^{i\mu\omega - (\sigma^2\omega^2)/2}$$

La función característica de $X_1 + X_2 + \ldots + X_n$ es entonces, según el teorema 3-12,

$$\phi(\omega) = \phi_1(\omega)\phi_2(\omega)\cdots\phi_n(\omega) = e^{in\mu\omega - (n\sigma^2\omega^2)/2}$$

dado que los X_j son independientes. Entonces, según el teorema 3-11, la función característica de

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

es

$$\phi_{\bar{X}}(\omega) = \phi\left(\frac{\omega}{n}\right) = e^{i\mu\omega - [(\sigma^2/n)\omega^2]/2}$$

Pero ésta es la función característica de una distribución normal con media μ y varianza σ^2/n, y el resultado deseado se deduce del teorema 3-13.

5.22. Demuestre el teorema 5-6, página 171.

Por definición, $(n-1)\hat{S}^2 = \sum_{j=1}^{n} (X_j - \bar{X})^2$. Entonces se deduce de (2), método 2 en el problema 5.20, que $V = V_1 + V_2$, donde

$$V = \sum_{j=1}^{n} \frac{(X_j - \mu)^2}{\sigma^2}, \qquad V_1 = \frac{(n-1)\hat{S}^2}{\sigma^2}, \qquad V_2 = \frac{(\bar{X} - \mu)^2}{\sigma^2/n}$$

Ahora según el teorema 4-3, página 125, V tiene distribución chi cuadrado con n grados de libertad [como se vio al remplazar X_j por $(X_j - \mu)/\sigma$]. También, según el problema 5.21, $\bar{X}$ tiene distribución normal con media μ y varianza σ^2/n. Por consiguiente, del teorema 4-3 con $v = 1$ y X_1 remplazada por $(\bar{X} - \mu)/\sqrt{\sigma^2/n}$, vemos que V_2 tiene distribución chi cuadrado con 1 grado de libertad. Se deduce del teorema 4-5, página 125, que si V_1 y V_2 son independientes, entonces V_1 tiene distribución chi cuadrado con $n - 1$ grados de libertad. Ya que se puede demostrar que V_1 y V_2 son realmente independientes, se obtiene el resultado requerido.

5.23. (a) Utilice el teorema 5-6 para determinar el número esperado de muestras en el problema 5.1 para el cual las varianzas muestrales son mayores que 7.2. (b) Verifique el resultado en (a) con el resultado real.

(a) Tenemos $n = 2$, $\sigma^2 = 10.8$ [del problema 5.1(b)]. Para $s_1^2 = 7.2$ tenemos

$$\frac{ns_1^2}{\sigma^2} = \frac{(2)(7.2)}{10.8} = 1.33$$

De acuerdo con el teorema 5-6, $\chi^2 = nS^2/\sigma^2 = 2S^2/10.8$ tiene distribución chi cuadrado con 1 grado de libertad. De la tabla en el apéndice E se deduce que

$$P(S^2 \geqq s_1^2) = P(\chi^2 \geqq 1.33) = 0.25$$

Por consiguiente, esperaríamos que alrededor del 25% de las muestras, o 6, tuvieran varianzas mayores que 7.2.

(b) Según el problema 5.18, encontramos por conteo que en realidad hay 6 varianzas mayores que 7.2, de manera que hay acuerdo.

CASO EN QUE SE DESCONOCE LA VARIANZA DE LA POBLACIÓN

5.24. Demuestre el teorema 5-7, página 171.

Sea $Y = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$, $Z = \dfrac{nS^2}{\sigma^2}$, $\nu = n - 1$. Entonces dado que las S_j tienen distribución normal con media μ y varianza σ^2, sabemos (problema 5.21) que $\bar{X}$ tiene distribución normal con media μ y varianza σ^2/n, de manera que y tiene distribución normal con media 0 y varianza 1. También, según el teorema 5-6, página 171, o problema 5.22, Z tiene distribución chi cuadrado con $v = n - 1$ grados de libertad. Más aún, se puede demostrar que Y y Z son independientes.

Se deduce del teorema 4-6, página 126, que

$$T = \frac{Y}{\sqrt{Z/\nu}} = \frac{\bar{X} - \mu}{S/\sqrt{n-1}} = \frac{\bar{X} - \mu}{\hat{S}/\sqrt{n}}$$

tiene distribución t con $n - 1$ grados de libertad.

5.25. De acuerdo con la tabla de la distribución t de Student para 1 grado de libertad (apéndice D), tenemos $P(-1.376 \leqq T \leqq 1.376) = 0.60$. Verifique si esto confirma los resultados obtenidos en el problema 5.1.

A partir de los valores de $\bar{X}$ en (1) de la página 167, y de los valores de S^2 en el problema 5.18(a), obtenemos los siguientes valores para $T = (\bar{X} - \mu)/(S/\sqrt{1})$:

$-\infty$	-7.0	-1.0	-0.33	0.11
-7.0	$-\infty$	-1.0	-0.20	0.25
-1.0	-1.0	$\cdots$	1.0	1.0
-0.33	-0.20	1.0	∞	2.33
0.11	0.25	1.0	2.33	∞

Hay 16 valores para los cuales $-1.376 \leqq T \leqq 1.376$, mientras que podríamos esperar $(0.60)(25) = 15$. Esto no es del todo malo considerando la pequeña cantidad de datos involucrados. Este método de muestreo fue de hecho la manera como Student originalmente obtuvo la distribución t.

DISTRIBUCIÓN MUESTRAL DE LAS RELACIONES DE VARIANZAS

5.26. Demuestre el teorema 5-8, página 172.

Llamemos las muestras de tamaños m y n por $X_1, \ldots, X_n$ y $Y_1, \ldots, Y_n$, respectivamente. Entonces las varianzas muestrales están dadas por

$$S_1^2 = \frac{1}{m}\sum_{j=1}^{m}(X_j - \bar{X})^2, \qquad S_2^2 = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \bar{Y})^2$$

donde $\bar{X}$ y $\bar{Y}$ son las medias muestrales.

Ahora, del teorema 5-6, página 171, sabemos que mS_1^2/σ_1^2 y nS_2^2/σ_2^2 tienen distribución chi cuadrado con $m-1$ y $n-1$ grados de libertad, respectivamente. Por consiguiente, del teorema 4-7, página 126, se deduce que

$$F = \frac{mS_1^2/(m-1)\sigma_1^2}{nS_2^2/(n-1)\sigma_2^2} = \frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2}$$

tiene distribución F con $m-1$, $n-1$ grados de libertad.

5.27. Se toman dos muestras de tamaño 8 y 10 de dos poblaciones, con distribución normal, con varianzas 20 y 36, respectivamente. Encuentre la probabilidad de que la varianza de la primera muestra sea más del doble de la varianza de la segunda.

Tenemos $m = 8$, $n = 10$, $\sigma_1^2 = 20$, $\sigma_2^2 = 36$. Por consiguiente,

$$F = \frac{8S_1^2/(7)(20)}{10S_2^2/(9)(36)} = 1.85\,\frac{S_1^2}{S_2^2}$$

El número de grados de libertad para el numerador y el denominador con $\nu_1 = m - 1 = 7$, $\nu_2 = n - 1 = 9$. Ahora, si S_1^2 es más del doble de S_2^2, es decir, $S_1^2 > 2S_2^2$, entonces $F > 3.70$. En relación con las tablas del apéndice F vemos que la probabilidad es menor que 0.05, pero más de 0.01. Para valores exactos necesitamos una tabla más extensa de la distribución F.

DISTRIBUCIONES DE FRECUENCIA

5.28. En la tabla 5-4 se registran los pesos de 40 estudiantes varones de la Universidad XYZ, aproximando a la libra más cercana. Construya la distribución de frecuencia.

Tabla 5-4

138	164	150	132	144	125	149	157
146	158	140	147	136	148	152	144
168	126	138	176	163	119	154	165
146	173	142	147	135	153	140	135
161	145	135	142	150	156	145	128

El peso mayor es 176 lb y el menor, 119 lb, así que el rango es $176 - 119 = 57$ lb.
Si se usan 5 intervalos de clase, el tamaño del intervalo de clase es $57/5 = 11$, aproximadamente.
Si se usan 20 intervalos de clase, el tamaño del intervalo de clase es $57/20 = 3$, aproximadamente.

Una selección adecuada para el intervalo de clase es 5 lb. También es conveniente escoger las marcas de clase como 120, 125, 130, 135, . . . libras. Por consiguiente, los intervalos de clase pueden tomarse como 118−122, 123−127, 128−132, . . . Con esta selección, los límites de clase son 117.5, 122.5, 127.5, . . ., los cuales no coinciden con los datos observados.

En la tabla 5-5 se muestra la distribución de frecuencia requerida. La columna central, denominada conteo, registro, se utiliza para tabular las frecuencias de clase de datos originales y generalmente se omite en la presentación final de la distribución de frecuencia.

Tabla 5-5

Peso (lb)	Cantidades	Frecuencia
118–122	/	1
123–127	//	2
128–132	//	2
133-137	////	4
138–142	///// /	6
143–147	///// ///	8
148–152	/////	5
153–157	////	4
158–162	//	2
163–167	///	3
168–172	/	1
173–177	//	2
	TOTAL	40

Tabla 5-6

Peso (lb)	Cantidades	Frecuencia
118–126	///	3
127–135	/////	5
136–144	///// ////	9
145-153	///// ///// //	12
154–162	/////	5
163–171	////	4
172–180	//	2
	TOTAL	40

Otra posibilidad.

Por supuesto que existen otras posibles distribuciones de frecuencia. Por ejemplo, la tabla 5-6 muestra una distribución de frecuencia con sólo 7 clases, en las cuales el intervalo de clase es 9 lb.

5.29. Construya un histograma y un polígono de frecuencia para los pesos de la distribución del problema 5.28.

El histograma y el polígono de frecuencia para cada uno de los dos casos considerados en el problema 5.28 están dados en las figuras 5-7 y 5-8. Observe que los centros de las bases de los rectángulos están ubicados en las marcas de clase.

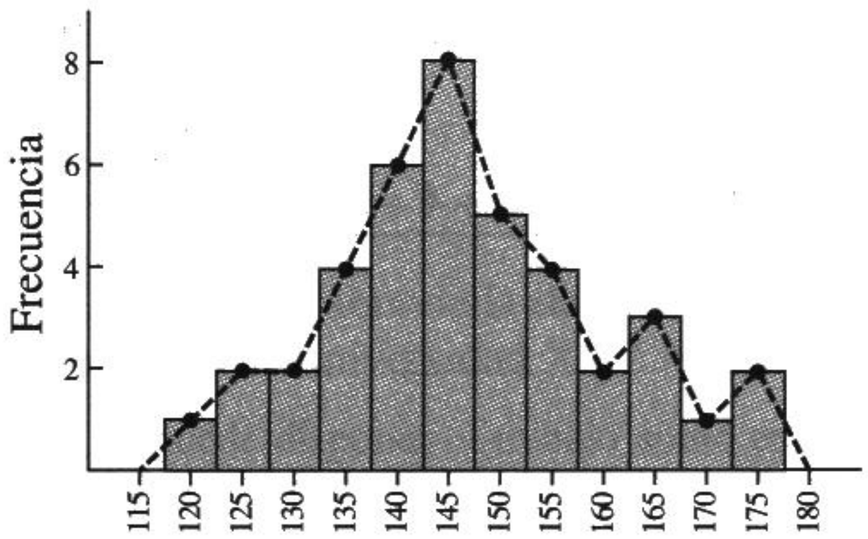


Fig. 5-7

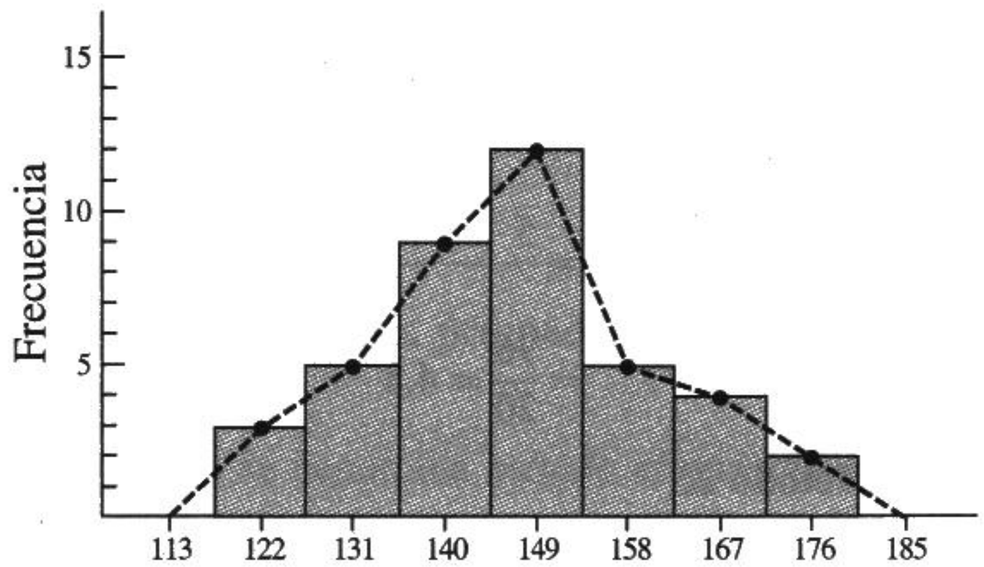


Fig. 5-8

5.30. Se lanzaron simultáneamente cinco monedas 1000 veces y en cada lanzamiento se observó el número de caras. En la tabla 5-7 se muestra el número de lanzamientos durante los cuales se obtuvieron 0, 1, 2, 3, 4 y 5 caras. Elabore una gráfica con los datos.

Los datos pueden mostrarse de manera gráfica bien sea como en la figura 5-9 o como en la figura 5-10.

Parece ser más adecuado usar la figura 5-9, ya que el número de caras no puede ser 1.5 ó 3.2. Ésta es una gráfica en forma de barras en donde las barras tienen ancho cero y se denomina *gráfico de vara*. Esto es especialmente útil cuando los datos son discretos.

La figura 5-10 muestra un histograma de los datos. Observe que el área total del histograma es el total de la frecuencia. 1000. como debe ser.

Tabla 5-7

Número de caras	Número de lanzamientos (frecuencia)
0	38
1	144
2	342
3	287
4	164
5	25
TOTAL	1000

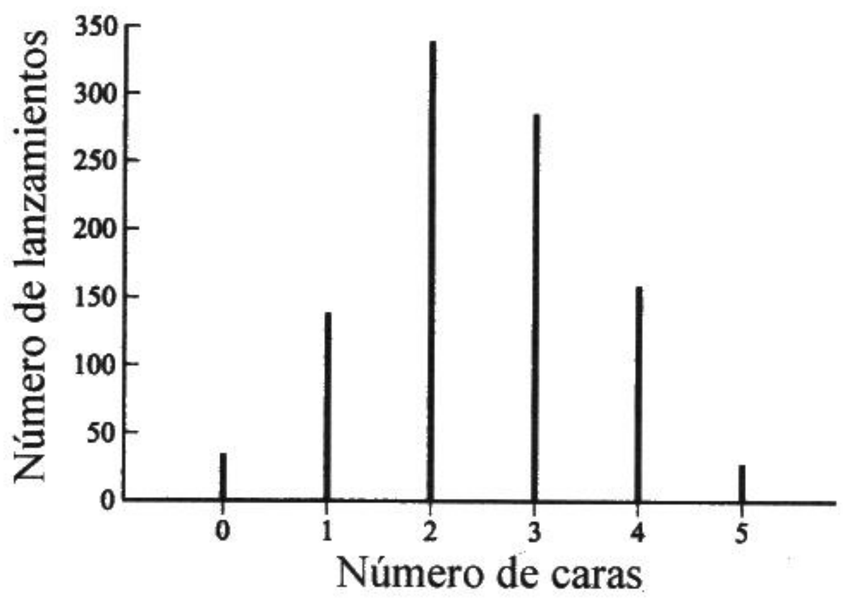


Fig. 5-9

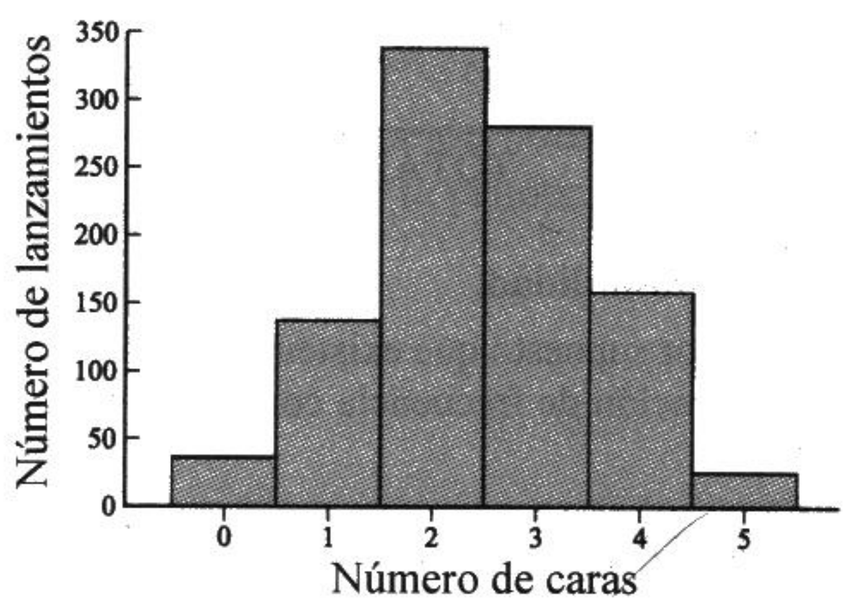


Fig. 5-10

CÁLCULO DE MEDIA, VARIANZA Y MOMENTOS PARA MUESTRAS

5.31. Encuentre la media aritmética de los números 5, 3, 6, 5, 4, 5, 2, 8, 6, 5, 4, 8, 3, 4, 5, 4, 8, 2, 5, 4.

Método 1.

$$\bar{x} = \frac{\sum x}{n} = \frac{5+3+6+5+4+5+2+8+6+5+4+8+3+4+5+4+8+2+5+4}{20}$$

$$= \frac{96}{20} = 4.8$$

Método 2.

Hay 6 cincos, 2 treses, 2 seises, 5 cuatros, 2 doses y 3 ochos. Entonces,

$$\bar{x} = \frac{\sum fx}{n} = \frac{(6)(5)+(2)(3)+(2)(6)+(5)(4)+(2)(2)+(3)(8)}{6+2+2+5+2+3} = \frac{96}{20} = 4.8$$

5.32. Se registran pesos promedio de 162, 148, 153 y 140 libras de cuatro grupos de estudiantes conformados por 15, 20, 10 y 18 individuos, respectivamente. Encuentre la media del peso de todos los estudiantes.

$$\bar{x} = \frac{\sum fx}{n} = \frac{(15)(162)+(20)(148)+(10)(153)+(18)(140)}{15+20+10+18} = 150\,\text{lb}$$

5.33. Use la distribución de frecuencia de estaturas de la tabla 5-2, página 173, para encontrar la estatura media de los 100 estudiantes varones de la Universidad XYZ.

El trabajo se resume en la tabla 5-8. Observe que todos los estudiantes que tienen estaturas 60-62 pulgadas, 63-65 pulgadas, etc. se consideran como si tuvieran estaturas de 61, 64 pulgadas, etc. Entonces, el problema se reduce a encontrar la estatura promedio de 100 estudiantes si 5 de ellos tuvieran estatura de 61 pulgadas, 18 de 64 pulgadas, etc.

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{\sum fx}{n} = \frac{6745}{100} = 67.45 \text{ pulgadas}$$

Tabla 5-8

Estatura (pulgadas)	Marca de clase (x)	Frecuencia (f)	fx
60-62	61	5	305
63–65	64	18	1152
66–68	67	42	2814
69–71	70	27	1890
72–74	73	8	584
		$n = \sum f = 100$	$\sum fx = 6745$

Los cálculos involucrados pueden ser tediosos, especialmente para los casos en los cuales los números son grandes y se presentan muchas clases. Se encuentran disponibles técnicas cortas para disminuir el trabajo en tales casos. Por ejemplo, *véase* el problema 5.35.

5.34. Derive la fórmula abreviada (*27*), página 174, para la media aritmética.

Digamos que la j-ésima marca de clase es x_j. Entonces la desviación de x_j de alguna marca de clase a especificada, la cual es $x_j - a$, será igual al intervalo de clase c multiplicado por algún entero u_j, a saber, $x_j - a = cu_j$ o $x_j = a + cu_j$ (también se puede escribir $x = a + cu$).

Entonces la media está dada por

$$\bar{x} = \frac{\sum f_j x_j}{n} = \frac{\sum f_j(a + cu_j)}{n} = \frac{a\sum f_j}{n} + c\frac{\sum f_j u_j}{n}$$

$$= a + c\frac{\sum f_j u_j}{n} = a + c\bar{u}$$

dado que $n = \sum f_j$.

5.35. Use la fórmula abreviada del problema 5.34 para encontrar la estatura promedio de 100 estudiantes varones de la Universidad XYZ (*véase* problema 5.33).

Se puede disponer el trabajo como en la tabla 5-9. El método se denomina *método abreviado* y se debe emplear siempre que sea posible.

$$\bar{x} = a + \left(\frac{\sum fu}{n}\right)c = 67 + \left(\frac{15}{100}\right)(3) = 67.45 \text{ pulgadas}$$

Tabla 5-9

	x	u	f	fu
	61	−2	5	−10
	64	−1	18	−18
$a \rightarrow$	67	0	42	0
	70	1	27	27
	73	2	8	16
			$n = 100$	$\sum fu = 15$

5.36. Encuentre (*a*) la varianza, (*b*) la desviación estándar de los números del problema 5.31.

(*a*) **Método 1.**

Tal como en el problema 5.31, tenemos $\bar{x} = 4.8$. Entonces,

$$s^2 = \frac{\sum (x - \bar{x})^2}{n} = \frac{(5 - 4.8)^2 + (3 - 4.8)^2 + (6 - 4.8)^2 + (5 - 4.8)^2 + \cdots + (4 - 4.8)^2}{20}$$

$$= \frac{59.20}{20} = 2.96$$

Método 2.

$$s^2 = \frac{\sum f(x - \bar{x})^2}{n} = \frac{6(5 - 4.8)^2 + 2(3 - 4.8)^2 + 2(6 - 4.8)^2 + 5(4 - 4.8)^2 + 3(8 - 4.8)^2}{20}$$

$$= \frac{59.20}{20} = 2.96$$

(*b*) De (*a*), $s^2 = 2.96$ y $s = \sqrt{2.96} = 1.72$.

5.37. Encuentre la desviación estándar de los pesos de los estudiantes del problema 5.32.

$$s^2 = \frac{\sum f(x - \bar{x})^2}{n} = \frac{15(162 - 150)^2 + 20(148 - 150)^2 + 10(153 - 150)^2 + 18(140 - 150)^2}{15 + 20 + 10 + 18}$$

$$= \frac{4130}{63} = 65.6 \text{ en unidades de libras al cuadrado o (libras)}^2$$

Entonces $s = \sqrt{65.6\,(\text{lb})^2} = \sqrt{65.6}\,\text{lb} = 8.10\,\text{lb}$, donde usamos el hecho de que esa unidad sigue las leyes básicas del álgebra.

5.38. Encuentre la desviación estándar de los pesos de los 100 estudiantes varones de la Universidad XYZ. *véase* el problema 5.33.

Del problema 5.33, $\bar{x} = 67.45$ pulgadas. El trabajo puede disponerse como en la tabla 5-10.

$$s = \sqrt{\frac{\sum f(x - \bar{x})^2}{n}} = \sqrt{\frac{852.7500}{100}} = \sqrt{8.5275} = 2.92 \text{ pulgadas}$$

Tabla 5-10

Estatura (pulgadas)	Marca de clase x	$x - \bar{x} = x - 67.45$	$(x - \bar{x})^2$	Frecuencia f	$f(x - \bar{x})^2$
60–62	61	−6.45	41.6025	5	208.0125
63–65	64	−3.45	11.9025	18	214.2450
66–68	67	−0.45	0.2025	42	8.5050
69–71	70	2.55	6.5025	27	175.5675
72–74	73	5.55	30.8025	8	246.4200
				$n = \sum f = 100$	$\sum f(x - \bar{x})^2 = 852.7500$

5.39. Derive la fórmula abreviada (*28*) para la varianza, página 174.

Tal como en el problema 5.34, tenemos $x_j = a + cu_j$ y

$$\bar{x} = a + c\,\frac{\sum f_j u_j}{n} = a + c\bar{u}$$

Entonces

$$s^2 = \frac{1}{n}\sum f_j(x_j - \bar{x})^2 = \frac{1}{n}\sum f_j(cu_j - c\bar{u})^2$$

$$= \frac{c^2}{n}\sum f_j(u_j - \bar{u})^2$$

$$= \frac{c^2}{n}\sum f_j(u_j^2 - 2u_j\bar{u} + \bar{u}^2)$$

$$= \frac{c^2}{n}\sum f_j u_j^2 - \frac{2\bar{u}c^2}{n}\sum f_j u_j + \frac{c^2}{n}\sum f_j\bar{u}^2$$

$$= c^2\,\frac{\sum f_j u_j^2}{n} - 2\bar{u}^2c^2 + c^2\bar{u}^2$$

$$= c^2\,\frac{\sum f_j u_j^2}{n} - c^2\left(\frac{\sum f_j u_j}{n}\right)^2$$

$$= c^2\left[\frac{\sum fu^2}{n} - \left(\frac{\sum fu}{n}\right)^2\right]$$

$$= c^2[\overline{u^2} - \bar{u}^2]$$

5.40. Use la fórmula abreviada del problema 5.39 para encontrar la desviación estándar de las estaturas del problema 5.33.

Podemos disponer el trabajo como en la tabla 5-11. Esto nos permite encontrar $\bar{x}$ como lo hicimos en el problema 5.35. A partir de la última columna tenemos entonces,

$$s^2 = c^2\left[\frac{\sum fu^2}{n} - \left(\frac{\sum fu^2}{n}\right)^2\right] = c^2(\overline{u^2} - \bar{u}^2)$$

$$= (3)^2\left[\frac{97}{100} - \left(\frac{15}{100}\right)^2\right] = 8.5275$$

y así $s = 2.92$ pulgadas

Tabla 5-11

	x	u	f	fu	fu^2
	61	−2	5	−10	20
	64	−1	18	−18	18
$a \longrightarrow$	67	0	42	0	0
	70	1	27	27	27
	73	2	8	8	32
			$n = \sum f = 100$	$\sum fu = 15$	$\sum fu^2 = 97$

5.41. Encuentre los primeros cuatro momentos alrededor de la media para la distribución de estaturas del problema 5.33.

Siguiendo con el método del problema 5.40, obtenemos la tabla 5-12. Luego, usando la notación de la página 175, tenemos

$$M_1' = \frac{\sum fu}{n} = 0.15 \qquad M_3' = \frac{\sum fu^3}{n} = 0.33$$

$$M_2' = \frac{\sum fu^2}{n} = 0.97 \qquad M_4' = \frac{\sum fu^4}{n} = 2.53$$

Tabla 5-12

x	u	f	fu	fu^2	fu^3	fu^4
61	−2	5	−10	20	−40	80
64	−1	18	−18	18	−18	18
67	0	42	0	0	0	0
70	1	27	27	27	27	27
73	2	8	16	32	64	128
		$n = \sum f = 100$	$\sum fu = 15$	$\sum fu^2 = 97$	$\sum fu^3 = 33$	$\sum fu^4 = 253$

y de (*32*),

$$m_1 = 0$$

$$m_2 = c^2(M_2' - M_1'^2) = 9[0.97 - (0.15)^2] = 8.5275$$

$$m_3 = c^3(M_3' - 3M_1'M_2' + 2M_1'^3) = 27[0.33 - 3(0.15)(0.97) + 2(0.15)^3] = -2.6932$$

$$m_4 = c^4(M_4' - 4M_1'M_3' + 6M_1'^2M_2' - 3M_1'^4)$$

$$= 81[2.53 - 4(0.15)(0.33) + 6(0.15)^2(0.97) - 3(0.15)^4] = 199.3759$$

5.42. Encuentre los coeficientes de (*a*) sesgo y (*b*) curtosis para la distribución de estaturas del problema 5.33.

(*a*) Del problema 5.41,

$$m_2 = s^2 = 8.5275 \qquad m_3 = -2.6932$$

Entonces,

$$\text{Coeficiente de sesgo} = a_3 = \frac{m_3}{s^3}$$

$$= \frac{-2.6932}{\sqrt{(8.5275)^3}} = -0.14$$

(*b*) Del problema 5.41,

$$m_4 = 199.3759 \qquad m_2 = s^2 = 8.5275$$

Entonces,

$$\text{Coeficiente de curtosis} = a_4 = \frac{m_4}{s^4}$$

$$= \frac{199.3759}{(8.5275)^2} = 2.74$$

De (*a*) vemos que la distribución está moderadamente sesgada hacia la izquierda. De (*b*) vemos que ligeramente tiene menos pico que la distribución normal (la cual tiene un coeficiente de curtosis = 3).

PROBLEMAS DIVERSOS

5.43. (*a*) Muestre cómo seleccionar 30 muestras aleatorias (con remplazo) cada una de 4 estudiantes de la tabla de estaturas de la página 173 usando números aleatorios. (*b*) Encuentre la media y la desviación estándar de la distribución muestral de medias en (*a*). (*c*) Compare los resultados de (*b*) con valores teóricos, y explique cualquier discrepancia.

(*a*) Use dos dígitos para enumerar cada uno de los 100 estudiantes: 00, 01, 02, ..., 99 (*véase* tabla 5-13). A continuación, enumere a los 5 estudiantes cuyas estaturas son 60-62 pulgadas como 00-04, los 18 estudiantes cuyas estaturas son 63-65 pulgadas como 05-22, etc. Cada número de estudiantes se llama *número muestral*.

Tabla 5-13

Estatura (pulgadas)	Frecuencia	Número muestral
60–62	5	00–04
63–65	18	05-22
66–68	42	23–64
69–71	27	65–91
72–74	8	92–99

Ahora obtenemos números muestrales de la tabla de números aleatorios (apéndice H). En la primera línea encontramos la secuencia 51, 77, 27, 46, 40, etc., números que tomamos como números muestrales aleatorios, cada uno de los cuales determina la estatura de un estudiante en particular. Por consiguiente, 51 corresponde a un estudiante cuya estatura está entre 66-68 pulgadas, la cual tomamos como 67 pulgadas (la marca de clase). De igual manera 77, 27, 46 representan las estaturas de 70, 67 y 67 pulgadas, respectivamente.

Por medio de este proceso llegamos a la tabla 5-14, que muestra los números muestrales obtenidos, sus estaturas correspondientes y la estatura media para cada una de las 30 muestras. Se debe mencionar que aunque entramos a la tabla de números aleatorios en el primer renglón, hubiéramos podido empezar en *cualquiera* y haber escogido cualquier patrón.

Tabla 5-14

Números muestrales obtenidos	Estaturas correspondientes	Estatura promedio	Números muestrales obtenidos	Estaturas correspondientes	Estatura promedio
1. 51, 77, 27, 46	67, 70, 67, 67	67.75	**16.** 11, 64, 55, 58	64, 67, 67, 67	66.25
2. 40, 42, 33, 12	67, 67, 67, 64	66.25	**17.** 70, 56, 97, 43	70, 67, 73, 67	69.25
3. 90, 44, 46, 62	70, 67, 67, 67	67.75	**18.** 74, 28, 93, 50	70, 67, 73, 67	69.25
4. 16, 28, 98, 93	64, 67, 73, 73	69.25	**19.** 79, 42, 71, 30	70, 67, 70, 67	68.50
5. 58, 20, 41, 86	67, 64, 67, 70	67.00	**20.** 58, 60, 21, 33	67, 67, 64, 67	66.25
6. 19, 64, 08, 70	64, 67, 64, 70	66.25	**21.** 75, 79, 74, 54	70, 70, 70, 67	69.25
7. 56, 24, 03, 32	67, 67, 61, 67	65.50	**22.** 06, 31, 04, 18	64, 67, 61, 64	64.00
8. 34, 91, 83, 58	67, 70, 70, 67	68.50	**23.** 67, 07, 12, 97	70, 64, 64, 73	67.75
9. 70, 65, 68, 21	70, 70, 70, 64	68.50	**24.** 31, 71, 69, 88	67, 70, 70, 70	69.25
10. 96, 02, 13, 87	73, 61, 64, 70	67.00	**25.** 11, 64, 21, 87	64, 67, 64, 70	66.25
11. 76, 10, 51, 08	70, 64, 67, 64	66.25	**26.** 03, 58, 57, 93	61, 67, 67, 73	67.00
12. 63, 97, 45, 39	67, 73, 67, 67	68.50	**27.** 53, 81, 93, 88	67, 70, 73, 70	70.00
13. 05, 81, 45, 93	64, 70, 67, 73	68.50	**28.** 23, 22, 96, 79	67, 64, 73, 70	68.50
14. 96, 01, 73, 52	73, 61, 70, 67	67.75	**29.** 98, 56, 59, 36	73, 67, 67, 67	68.50
15. 07, 82, 54, 24	64, 70, 67, 67	67.00	**30.** 08, 15, 08, 84	64, 64, 64, 70	65.50

(*b*) La tabla 5-15 muestra la distribución de frecuencia de la media muestral de la estatura obtenida en (*a*). Ésta es una *distribución muestral de medias*. La media y la desviación estándar se obtienen de manera usual mediante los métodos abreviados que ya describimos.

$$\text{Media} = a + c\bar{u} = a + \frac{c\sum fu}{n} = 67.00 + \frac{(0.75)(23)}{30} = 67.58 \text{ pulgadas}$$

$$\text{Desviación estándar} = c\sqrt{\bar{u^2} - \bar{u}^2} = c\sqrt{\frac{\sum fu^2}{n} - \left(\frac{\sum fu}{n}\right)^2}$$

$$= (0.75)\sqrt{\frac{123}{30} - \left(\frac{23}{30}\right)^2} = 1.41 \text{ pulgadas}$$

Tabla 5-15

Media muestral	Conteo	f	u	fu	fu^2
64.00	/	1	−4	−4	16
64.75		0	−3	0	0
65.50	//	2	−2	−4	8
66.25	///// /	6	−1	−6	6
$a \longrightarrow$ 67.00	////	4	0	0	0
67.75	////	4	1	4	4
68.50	///// //	7	2	14	28
69.25	/////	5	3	15	45
70.00	/	1	4	4	16
		$\sum f = n = 30$		$\sum fu = 23$	$\sum fu^2 = 123$

(*c*) La media teórica de la distribución muestral de medias, dada por $\mu_{\bar{X}}$, es igual a la media de la población μ, que es 67.45 pulgadas (*véase* problema 5.33), lo cual está muy de acuerdo con el valor 67.58 pulgadas de la parte (*b*).

La desviación estándar teórica (error estándar) de la distribución muestral de medias, dada por $\sigma_{\bar{X}}$, es igual a $\sigma/\sqrt{n}$, donde la desviación estándar de la población $\sigma = 2.92$ pulgadas (*véase* problema 5.40) y el tamaño de la muestra $n = 4$. Dado que $\sigma/\sqrt{n} = 2.92/\sqrt{4} = 1.46$ pulgadas, podemos decir que hay cierto acuerdo con el valor 1.41 pulgadas de la parte (*b*). Las discrepancias se deben al hecho de que solamente se seleccionaron 30 muestras y el tamaño de la muestra fue pequeño.

5.44. La desviación estándar de los pesos de una población muy grande de estudiantes es 10.0 lb. Se tomaron muestras de esta población de 200 estudiantes cada una, y se calcularon las desviaciones estándar de los pesos de cada muestra. Encuentre (*a*) la media y (*b*) la desviación estándar de la distribución muestral de las desviaciones estándar.

Podemos considerar que las muestras se tomaron de una población infinita o de una población finita con remplazo. De la tabla 5-1, página 172, tenemos

(*a*) $$\mu_S = \sigma = 10.0\,\text{lb}$$

(*b*) $$\sigma_S = \frac{\sigma}{\sqrt{2n}} = \frac{10}{\sqrt{400}} = 0.50\,\text{lb}$$

5.45. ¿Qué porcentaje de las muestras del problema 5.44 tendrá desviaciones estándar (*a*) mayores que 11.0 lb, (*b*) menores que 8.8 lb?

La distribución muestral de las desviaciones estándar es aproximadamente normal con media 10.0 lb y desviación estándar 0.50 lb.

(*a*) 11.0 lb en unidades estándar $= (11.0 - 10.0)/0.50 = 2.0$. El área bajo la curva normal a la derecha de $z = 2.0$ es $(0.5 - 0.4772) = 0.0228$; así, el porcentaje requerido es 2.3%.

(*b*) 8.8 lb en unidades estándar $= (8.8 - 10.0)/0.50 = -2.4$. El área bajo la curva normal a la izquierda de $z = -2.4$ es $(0.5 - 0.4918) = 0.0082$; así, el porcentaje requerido es 0.8%.

5.46. De una población continua se obtiene al azar una muestra de 6 observaciones. ¿Cuál es la probabilidad de que las últimas 2 observaciones sean menores que las primeras 4?

Supongamos que la población tiene función de densidad $f(x)$. La probabilidad de que 3 de las primeras 4 observaciones sean mayores que μ mientras la cuarta observación se encuentra entre u y $u + du$ está dada por

(*1*) $$ {}_4C_3\left[\int_u^\infty f(x)\,dx\right]^3 f(u)\,du$$

La probabilidad de que las últimas 2 observaciones sean menores que u (y por lo tanto menores que las primeras 4) está dada por

$$(2) \qquad \left[\int_{-\infty}^{u} f(x)\,dx\right]^2$$

Entonces, la probabilidad de que las primeras 4 sean mayores que u y las últimas 2 sean menores que u, es el producto de (*1*) y (*2*), esto es,

$$(3) \qquad {}_4C_3\left[\int_{u}^{\infty} f(x)\,dx\right]^3 f(u)\,du\left[\int_{-\infty}^{u} f(x)\,dx\right]^2$$

Dado que u puede tomar valores entre $-\infty$ e ∞, la probabilidad total de que las últimas 2 observaciones sean menores que las primeras 4 es la integral de (*3*) desde $-\infty$ hasta ∞, así

$$(4) \qquad {}_4C_3\int_{-\infty}^{\infty}\left[\int_{u}^{\infty} f(x)\,dx\right]^3\left[\int_{-\infty}^{u} f(x)\,dx\right]^2 f(u)\,du$$

Para calcular esto, digamos que

$$(5) \qquad v=\int_{-\infty}^{u} f(x)\,dx$$

Entonces

$$(6) \qquad dv=f(u)\,du \qquad\qquad 1-v=\int_{u}^{\infty} f(x)\,dx$$

Cuando $u=\infty$, $v=1$, y cuando $u=-\infty$, $v=0$. Por consiguiente, (*4*) se convierte en

$${}_4C_3\int_{0}^{1} v^2(1-v)^3\,dv = 4\,\frac{\Gamma(3)\Gamma(4)}{\Gamma(7)}=\frac{1}{15}$$

lo cual es la probabilidad requerida. Es interesante anotar que la probabilidad no depende de la distribución de probabilidad $f(x)$. Esto es un ejemplo de la estadística no paramétrica, dado que no es necesario conocer los parámetros de la población.

Otro método.

Llamemos las observaciones $x_1, x_2, \ldots, x_6$. Ya que la población es continua, podemos suponer que todas las x_i son diferentes. Hay 6! maneras de disponer los subíndices 1, 2 ..., 6 en orden creciente. De las 6!, exactamente $4! \times 2!$ arreglos tendrán x_1, x_2, x_3, x_4 como las 4 observaciones más pequeñas y x_5, x_6 como las últimas 2 observaciones más grandes. La probabilidad requerida es, entonces,

$$\frac{4!\times 2!}{6!}=\frac{1}{15}$$

5.47. Sea $\{X_1, X_2, \ldots, X_n\}$ una muestra aleatoria de tamaño n que se obtuvo sin remplazo de una población finita de tamaño N. Demuestre que si la media y la varianza de la población son μ y σ^2, entonces (*a*) $E(X_j)=\mu$, (*b*) $\mathrm{Cov}(X_j, X_k)=-\sigma^2/(N-1)$.

Supongamos que la población consta del conjunto de números $\{\alpha_1, \alpha_2, \ldots, \alpha_N\}$, donde los α no necesariamente son diferentes. *Un procedimiento de muestreo aleatorio es aquel en el cual cada selección de n fuera de N α tiene la misma probabilidad de ocurrencia (a saber, $1/{}_NC_n$). Esto significa que las X_j tienen distribución idéntica.*

$$X_j=\begin{cases}\alpha_1 & \text{prob. } 1/N\\ \alpha_2 & \text{prob. } 1/N\\ \vdots & \\ \alpha_N & \text{prob. } 1/N\end{cases} \qquad (j=1,2,\ldots,n)$$

Sin embargo, no son mutuamente independientes. Es más, cuando $j \neq k$, la distribución conjunta de X_j y X_k está dada por

$$P(X_j = \alpha_\lambda, X_k = \alpha_\nu) = P(X_j = \alpha_\lambda)P(X_k = \alpha_\nu \mid X_j = \alpha_\lambda)$$
$$= \frac{1}{N} P(X_k = \alpha_\nu \mid X_j = \alpha_\lambda)$$
$$= \begin{cases} \frac{1}{N}\left(\frac{1}{N-1}\right) & \lambda \neq \nu \\ 0 & \lambda = \nu \end{cases}$$

donde λ y ν varían de 1 a N.

(*a*)
$$E(X_j) = \sum_{\lambda=1}^{N} \alpha_\lambda P(X_j = \alpha_\lambda) = \frac{1}{N} \sum_{\lambda=1}^{N} \alpha_\lambda = \mu$$

(*b*)
$$\text{Cov}\,(X_j, X_k) = E[X_j - \mu)(X_k - \mu)]$$
$$= \sum_{\lambda=1}^{N} \sum_{\nu=1}^{N} (\alpha_\lambda - \mu)(\alpha_\nu - \mu)P(X_j = \alpha_\lambda, X_k = \alpha_\nu)$$
$$= \frac{1}{N}\left(\frac{1}{N-1}\right) \sum_{\lambda \neq \nu = 1}^{N} (\alpha_\lambda - \mu)(\alpha_\nu - \mu)$$

donde la última suma contiene un total de $N(N-1)$ términos correspondientes a todos los pares posibles de λ y ν desiguales.

Ahora, según el álgebra elemental,

$$[(\alpha_1 - \mu) + (\alpha_2 - \mu) + \cdots + (\alpha_N - \mu)]^2 = \sum_{\lambda=1}^{N} (\alpha_\lambda - \mu)^2 + \sum_{\lambda \neq \nu = 1}^{N} (\alpha_\lambda - \mu)(\alpha_\nu - \mu)$$

En esta ecuación, el lado izquierdo es cero, ya que por definición

$$\alpha_1 + \alpha_2 + \cdots + \alpha_N = N\mu$$

y la primera suma del lado derecho es, por definición, igual a $N\sigma^2$. Así que

$$\sum_{\lambda \neq \nu = 1}^{N} (\alpha_\lambda - \mu)(\alpha_\nu - \mu) = -N\sigma^2$$

y
$$\text{Cov}\,(X_j, X_k) = \frac{1}{N}\left(\frac{1}{N-1}\right)(-N\sigma^2) = -\frac{\sigma^2}{N-1}$$

5.48. Demuestre que (*a*) la media, (*b*) la varianza de la media muestral del problema 5.47 están dadas, respectivamente, por

$$\mu_{\bar{X}} = \mu \qquad \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

(*a*)
$$E(\bar{X}) = E\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n}[E(X_1) + \cdots + E(X_n)]$$
$$= \frac{1}{n}(\mu + \cdots + \mu) = \mu$$

donde usamos el problema 5.47(*a*).

(*b*) Usando los teoremas 3-5 y 3-16 (generalizado), y el problema 5.47, obtenemos

$$\text{Var}(\bar{X}) = \frac{1}{n^2}\text{Var}\left(\sum_{j=1}^{n} X_j\right) = \frac{1}{n^2}\left[\sum_{j=1}^{n}\text{Var}(X_j) + \sum_{j\neq k=1}^{n}\text{Cov}(X_j, X_k)\right]$$

$$= \frac{1}{n^2}\left[n\sigma^2 + n(n-1)\left(-\frac{\sigma^2}{N-1}\right)\right]$$

$$= \frac{\sigma^2}{n}\left[1 - \frac{n-1}{N-1}\right] = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

Problemas suplementarios

DISTRIBUCIÓN MUESTRAL DE MEDIAS

5.49. Una población consta de los cuatro números 3, 7, 11, 15. Considere todas las muestras posibles de tamaño dos que se pueden obtener con remplazo de esta población. Encuentre (*a*) la media de la población, (*b*) la desviación estándar de la población, (*c*) la media de la distribución muestral de medias, (*d*) la desviación estándar de la distribución muestral de medias. Verifique (*c*) y (*d*) directamente a partir de (*a*) y (*b*) usando fórmulas apropiadas.

5.50. Resuelva el problema 5.49 si se toman muestras sin remplazo.

5.51. Los pesos de 1500 balineras tienen distribución normal con media 22.40 onzas y desviación estándar 0.048 onzas. Si de esta población se toman 300 muestras aleatorias de tamaño 36, determine la media esperada y la desviación estándar de la distribución muestral de medias si el muestreo se hizo (*a*) con remplazo, (*b*) sin remplazo.

5.52. Resuelva el problema 5.51 si la población consta de 72 balineras.

5.53. ¿Cuántas muestras aleatorias del problema 5.51 tendrán su media (*a*) entre 22.39 y 22.41 oz, (*b*) mayor que 22.42 oz, (*c*) menor que 22.37 oz, (*d*) menor que 22.38 o mayor que 22.41 oz?

5.54. Ciertos tubos fabricados por una compañía tienen tiempo de vida media de 800 horas y desviación estándar de 60 horas. Encuentre la probabilidad de que una muestra aleatoria de 16 tubos tomados del grupo tengan tiempo de vida media (*a*) entre 790 y 810 horas, (*b*) menos de 785 horas, (*c*) más de 820 horas, (*d*) entre 770 y 830 horas.

5.55. Desarrolle el problema 5.54 si se toma una muestra aleatoria de 64 tubos. Explique la diferencia.

5.56. El peso de los paquetes que recibe una bodega tiene una media de 300 lb y una desviación estándar de 50 lb. ¿Cuál es la probabilidad de que un ascensor cargado con 25 paquetes, recibidos al azar, exceda su límite de seguridad, considerado como 8200 lb?

DISTRIBUCIÓN MUESTRAL DE PROPORCIONES

5.57. Encuentre la probabilidad de que de los siguientes 200 bebés que nazcan (*a*) menos del 40% sean niños, (*b*) entre 43% y 57% sean niñas, (*c*) más de 54% sean niños. Suponga que la probabilidad de nacimiento de niña y niño es igual.

5.58. De 1000 muestras cada una de 200 infantes, ¿en cuántas esperaría usted encontrar que (*a*) menos de 40% son niños, (*b*) entre 40% y 60% son niñas, (*c*) 53% o más son niñas?

5.59. Desarrolle el problema 5.57 si en lugar de considerar muestras de 200, se toman de 100, y explique la diferencia en los resultados.

5.60. Una urna contiene 80 canicas de las cuales 60% son rojas y 40% son blancas. De 50 muestras de 20 canicas, cada una seleccionada con remplazo, ¿cuántas muestras pueden esperarse que consten de (*a*) número igual de canicas rojas y blancas, (*b*) 12 canicas rojas y 8 blancas, (*c*) 8 canicas rojas y 12 blancas, (*d*) 10 o más canicas blancas?

5.61. Diseñe un experimento destinado a ilustrar los resultados del problema 5.60. En lugar de canicas rojas y blancas, usted puede usar trozos de papel sobre los cuales puede escribir *R* o *B* en las proporciones correspondientes. ¿Qué error puede introducir si utiliza dos conjuntos de canicas?

5.62. Un fabricante despacha 1000 lotes, cada uno de 100 bombillas eléctricas. Si normalmente el 5% de las bombillas es defectuoso, ¿en cuántos lotes esperaría usted encontrar (*a*) menos de 90 bombillas buenas, (*b*) 98 o más bombillas buenas?

DISTRIBUCIÓN MUESTRAL DE DIFERENCIAS Y SUMAS

5.63. *A* y *B* fabrican dos tipos de cables cuya resistencia promedio al rompimiento es de 4000 y 5000 lb con desviaciones estándar de 300 y 200 lb, respectivamente. Si se prueban 100 cables de la marca *A* y 50 cables de la marca *B*, ¿cuál es la probabilidad de que la resistencia promedio al rompimiento de *B* sea (*a*) al menos 600 lb mayor que la de *A*, (*b*) al menos 450 lb mayor que la de *A*?

5.64. ¿Cuáles son las probabilidades en el problema 5.63 si se prueban 100 cables de ambas marcas? Explique las diferencias.

5.65. El puntaje promedio de ciertos estudiantes en una prueba de aptitud es 72 puntos con desviación estándar de 8 puntos. ¿Cuál es la probabilidad de que dos grupos de 28 y 36 estudiantes, respectivamente, difieran en sus puntajes promedio por (*a*) 3 puntos o más, (*b*) 6 puntos o más, (*c*) entre 2 y 5 puntos?

5.66. Una urna contiene 60 canicas rojas y 40 blancas. Se sacan, con remplazo, dos conjuntos de 30 canicas cada uno y se anotan sus colores. ¿Cuál es la probabilidad de que los dos conjuntos difieran por 8 o más canicas rojas?

5.67. Resuelva el problema 5.66 si el muestreo para obtener cada conjunto se lleva a cabo sin remplazo.

5.68. Los resultados de las elecciones mostraron que cierto candidato obtuvo el 65% de los votos. Encuentre la probabilidad de que dos muestras aleatorias, cada una de 200 votos, indiquen una diferencia mayor del 10% en la proporción de votos para ese candidato.

5.69. Si U_1 y U_2 son los conjuntos de números del problema 5.12, verifique que (*a*) $\mu_{U_1+U_2} = \mu_{U_1} + \mu_{U_2}$, $\sigma_{U_1+U_2} = \sqrt{\sigma_{U_1}^2 + \sigma_{U_2}^2}$.

5.70. Se obtienen tres pesos de 20.48, 35.97 y 62.34 lb con desviación estándar 0.21, 0.46 y 0.54 lb, respectivamente. Encuentre (*a*) la media y (*b*) la desviación estándar de la suma de los pesos.

5.71. El voltaje de una batería tiene casi distribución normal con media 15.0 voltios y desviación estándar 0.2 voltios. ¿Cuál es la probabilidad de que cuatro de las baterías conectadas en serie tengan un voltaje combinado de 60.8 o más voltios?

DISTRIBUCIÓN MUESTRAL DE VARIANZAS

5.72. En relación con el problema 5.49, encuentre (*a*) la media de la distribución muestral de varianzas, (*b*) el error estándar de las varianzas.

5.73. Desarrolle el problema 5.72 si se toman muestras sin remplazo.

5.74. Una población normal tiene varianza 15. Si de esta población se obtienen muestras de tamaño 5, ¿qué porcentaje se puede esperar que tenga varianza (*a*) menor que 10, (*b*) mayor que 20, (*c*) entre 5 y 10?

5.75. Se encuentra que el tiempo de vida media de los tubos de televisión fabricados por una compañía tiene distribución normal con media de 2000 horas y desviación estándar de 60 horas. Si se seleccionan al azar 10 tubos, encuentre la probabilidad de que la desviación estándar muestral (*a*) no exceda 50 horas, (*b*) esté entre 50 y 70 horas.

CASO EN QUE SE DESCONOCE LA VARIANZA DE LA POBLACIÓN

5.76. Según la tabla de la distribución *t* de Student para 1 grado de libertad (apéndice D), tenemos $P(-1 \leq T \leq 1) = 0.50$. Verifique si los resultados del problema 5.1 se confirman con este valor, y explique cualquier diferencia.

5.77. Verifique si se confirman los resultados del problema 5.49 usando (*a*) $P(-1 \leq T \leq 1) = 0.50$, (*b*) $P(-1.376 \leq T \leq 1.376) = 0.60$, donde T tiene distribución *t* de Student con $v = 1$.

5.78. Explique cómo puede usar el teorema 5-7, página 171, para diseñar una tabla de la distribución *t* de Student como la del apéndice D.

DISTRIBUCIÓN MUESTRAL DE LAS RELACIONES DE VARIANZAS

5.79. De una población con distribución normal se obtienen dos muestras de tamaño 4 y 8. ¿Existe la probabilidad de que una varianza sea mayor que la otra 1.5 veces, mayor que 0.05, entre 0.05 y 0.01, o menor que 0.01?

5.80. Dos compañías, *A* y *B*, fabrican bombillas eléctricas. El tiempo de vida de ambas tiene distribución normal. Las bombillas de *A* tienen desviación estándar de 40 horas, mientras que el tiempo de vida de las bombillas de *B* tiene desviación estándar de 50 horas. Se toma una muestra de 8 bombillas de *A* y 16 bombillas de *B*. Determine la probabilidad de que la varianza de la primera muestra sea mayor que (*a*) el doble, (*b*) 1.2 veces la segunda.

5.81. Desarrolle el problema 5.80 si la desviación estándar de los tiempos de vida son (*a*) ambos de 40 horas, (*b*) ambos de 50 horas.

DISTRIBUCIÓN DE FRECUENCIA

5.82. La tabla 5-16 muestra la distribución de frecuencia de los tiempos de vida de 400 tubos que se probaron en la compañía L & M. Con relación a esta tabla determine

(*a*) el límite superior de la quinta clase,

(*b*) el límite inferior de la octava clase,

(*c*) la marca de clase de la séptima clase,

Tabla 5-16

Tiempo de vida (horas)	Número de tubos
300–399	14
400–499	46
500–599	58
600–699	76
700–799	68
800–899	62
900–999	48
1000–1099	22
1100–1199	6
TOTAL	400

(*d*) los límites de clase de la última clase,
(*e*) el tamaño del intervalo de clase,
(*f*) la frecuencia de la cuarta clase
(*g*) la frecuencia relativa de la sexta clase,
(*h*) el porcentaje de tubos cuyos tiempos de vida no exceden 600 horas,
(*i*) el porcentaje de tubos cuyos tiempos de vida son mayores o iguales que 900 horas,
(*j*) el porcentaje de tubos cuyos tiempos de vida son al menos 500, pero menores que 1000 horas.

5.83. Construya (*a*) un histograma, (*b*) un polígono de frecuencia, correspondiente a la distribución de frecuencia del problema 5.82.

5.84. Construya, para los datos del problema 5.82, (*a*) la distribución de frecuencia relativa o de porcentaje, (*b*) un histograma de frecuencia relativa, (*c*) un polígono de frecuencia relativa.

5.85. Estime el porcentaje de tubos del problema 5.82 con tiempos de vida de (*a*) menos de 560 horas, (*b*) 970 horas o más, (*c*) entre 620 y 890 horas.

5.86. Los diámetros internos de los empaques producidos por una fábrica se pueden medir con aproximación de una milésima de centímetro. Si la marca de clase de la distribución de frecuencia de estos diámetros está dada en centímetros por 0.321, 0.324, 0.327, 0.330, 0.333 y 0.336, encuentre (*a*) el tamaño del intervalo de clase, (*b*) los límites reales de clase, (*c*) los límites de clase.

5.87. La tabla 5-17 muestra los diámetros en centímetros de una muestra de 60 balineras fabricadas por una compañía. Construya la distribución de frecuencia de los diámetros usando intervalos de clase apropiados.

Tabla 5-17

0.738	0.729	0.743	0.740	0.736	0.741	0.735	0.731	0.726	0.737
0.728	0.737	0.736	0.735	0.724	0.733	0.742	0.736	0.739	0.735
0.745	0.736	0.742	0.740	0.728	0.738	0.725	0.733	0.734	0.732
0.733	0.730	0.732	0.730	0.739	0.734	0.738	0.739	0.727	0.735
0.735	0.732	0.735	0.727	0.734	0.732	0.736	0.741	0.736	0.744
0.732	0.737	0.731	0.746	0.735	0.735	0.729	0.734	0.730	0.740

5.88. Para los datos del problema 5.87 elabore (*a*) un histograma, (*b*) un polígono de frecuencia, (*c*) una distribución de frecuencia relativa, (*d*) un histograma de frecuencia relativa, (*e*) un polígono de frecuencia relativa.

5.89. A partir de los resultados del problema 5-88, determine el porcentaje de balineras con diámetros (*a*) que excedan 0.732 cm, (*b*) no mayor de 0.736 cm, (*c*) entre 0.730 y 0.738 centímetros. Compare sus resultados con aquéllos obtenidos directamente a partir de los datos originales de la tabla 5-17.

5.90. Desarrolle el problema 5.88 para los datos del problema 5.82.

CÁLCULO DE MEDIA, DESVIACIÓN ESTÁNDAR Y MOMENTOS PARA MUESTRAS

5.91. Un estudiante recibe notas de 85, 76, 93, 82 y 96 en cinco materias. Determine la media aritmética de las notas.

5.92. Un psicólogo midió los tiempos de reacción de un individuo a cierto estímulo y obtuvo 0.53, 0.46, 0.50, 0.49, 0.52, 0.53, 0.44 y 0.55 segundos. Determine el tiempo medio de reacción del individuo a los estímulos.

5.93. Un conjunto de números consta de 6 seises, 7 sietes, 8 ochos, 9 nueves y 10 dieces. ¿Cuál es la media aritmética de los números?

5.94. Las notas de un estudiante en laboratorio, lectura y física fueron 71, 78 y 89, respectivamente. (*a*) Si las notas ponderadas para estas materias son 2, 4 y 5, respectivamente, ¿cuál es una nota promedio apropiada? (*b*) ¿Cuál es la nota promedio si se usan notas ponderadas iguales?

5.95. Tres profesores de economía reportaron notas promedio de exámenes de 79, 74 y 82 en sus clases, las cuales tenían 32, 25 y 17 estudiantes, respectivamente. Determine la nota media para todas las clases.

5.96. El salario promedio anual que se pagó a los empleados de una compañía fue US$5000. El salario anual promedio que se pagó a los hombres y a las mujeres de esa compañía fue US$5200 y US$4200, respectivamente. Determine el porcentaje de hombres y mujeres que emplea la compañía.

5.97. La tabla 5-18 muestra la distribución de la carga máxima en toneladas cortas (1 tonelada corta = 2000 lb) que soportan algunos cables producidos por una compañía. Determine la carga máxima promedio usando (*a*) el "método largo", (*b*) el método abreviado.

Tabla 5-18

Carga máxima (toneladas cortas)	Número de cables
9–9.7	2
9.8–10.2	5
10.3–10.7	12
10.8–11.2	17
11.3–11.7	14
11.8–12.2	6
12.3–12.7	3
12.8–13.2	1
TOTAL	60

Tabla 5-19

x	462	480	498	516	534	552	570	588	606	624
f	98	75	56	42	30	21	15	11	6	2

5.98. Para los datos de la tabla 5-19 encuentre $\bar{x}$ usando (*a*) el método largo, (*b*) el método abreviado.

5.99. La tabla 5-20 muestra la distribución del diámetro de las cabezas de remaches fabricadas por una compañía. Calcule la media del diámetro.

Tabla 5-20

Diámetro (centímetros)	Frecuencia
0.7247–0.7249	2
0.7250–0.7252	6
0.7253–0.7255	8
0.7256–0.7258	15
0.7259–0.7261	42
0.7262–0.7264	68
0.7265–0.7267	49
0.7268–0.7270	25
0.7271–0.7273	18
0.7274–0.7276	12
0.7277–0.7279	4
0.7280–0.7282	1
TOTAL	250

5.100. Calcule la media de los datos de la tabla 5-21.

Tabla 5-21

Clase	Frecuencia
10– menos de 15	3
15– menos de 20	7
20– menos de 25	16
25– menos de 30	12
30– menos de 35	9
35– menos de 40	5
40– menos de 45	2
TOTAL	54

5.101. Encuentre la desviación estándar de los números:

(*a*) 3, 6, 2, 1, 7, 5; (*b*) 3.2, 4.6, 2.8, 5.2, 4.4; (*c*) 0, 0, 0, 0, 0, 1, 1, 1.

5.102. (*a*) Agregando 5 a cada uno de los números en el conjunto 3, 6, 2, 1, 7, 5 obtenemos el conjunto 8, 11, 7, 6, 12, 10. Demuestre que ambos conjuntos tienen la misma desviación estándar, pero medias diferentes. ¿Cómo están relacionadas las medias?

(*b*) Multiplicando cada uno de los números 3, 6, 2, 1, 7, 5 por 2 y luego sumando 5, obtenemos el conjunto 11, 17, 9, 7, 19, 15. ¿Cuál es la relación entre las desviaciones estándar y entre las medias de los dos conjuntos?

(*c*) ¿Cuáles de las propiedades de la media y de la desviación estándar se ilustran en el conjunto particular de números en (*a*) y (*b*)?

5.103. Encuentre la desviación estándar del conjunto de números en la progresión aritmética 4, 10, 16, 22, ..., 154.

5.104. Encuentre las desviaciones estándar para las distribuciones de (*a*) el problema 5-97, (*b*) el problema 5.98.

5.105. Encuentre (*a*) la media y (*b*) la desviación estándar para la distribución del problema 5.30, y explique la significancia de los resultados obtenidos.

5.106. (*a*) Encuentre la desviación estándar s del diámetro de las cabezas de remaches del problema 5.99. (*b*) ¿Qué porcentaje de los diámetros de las cabezas de remaches están en$(\bar{x} \pm s)$, $(\bar{x} \pm 2s)$, $(\bar{x} \pm 3s)$? (*c*) Compare los porcentajes en (*b*) con los que se esperarían teóricamente si la distribución fuera normal, y explique las diferencias observadas.

5.107. (*a*) Encuentre la media y la desviación estándar de los datos del problema 5.28.

(*b*) Construya la distribución de frecuencia de los datos y encuentre la desviación estándar.

(*c*) Compare el resultado de (*b*) con los de (*a*).

5.108. Desarrolle el problema 5.107 para los datos del problema 5.87.

5.109. (*a*) De un total de n números, la fracción p está compuesta de unos, mientras que la fracción $q = 1 - p$ está compuesta de ceros. Demuestre que la desviación estándar del conjunto de números es $\sqrt{pq}$. (*b*) Aplique el resultado de (*a*) al problema 5.101(*c*).

5.110. Encuentre (*a*) el primero, (*b*) el segundo, (*c*) el tercero, (*d*) el cuarto momento alrededor del origen para el conjunto de números 4, 7, 5, 9, 8, 3, 6.

5.111. Encuentre (*a*) el primero, (*b*) el segundo, (*c*) el tercero, (*d*) el cuarto momento alrededor de la media del conjunto de números del problema 5.110.

5.112. Encuentre (*a*) el primero, (*b*) el segundo, (*c*) el tercero, (*d*) el cuarto momento alrededor del número 7 para el conjunto de números del problema 5.110.

5.113. Usando los resultados de los problemas 5.110 y 5.111, verifique las siguientes relaciones entre los momentos: (*a*) $m_2 = m_2' - m_1'^2$, (*b*) $m_3 = m_3' - 3m_1'm_2' + 2m_1'^3$, (*c*) $m_4 = m_4' - 4m_1'm_3' + 6m_1'^2m_2' - 3m_1'^4$.

5.114. Encuentre los primeros cuatro momentos alrededor de la media del conjunto de números de la progresión aritmética 2, 5, 8, 11, 14, 17.

5.115. Si el primer momento alrededor del número 2 es igual a 5, ¿cuál es la media?

5.116. Si los primeros cuatro momentos de un conjunto de números alrededor del número 3 son iguales a -2, 10, -25 y 50, determine los momentos correspondientes (*a*) alrededor de la media, (*b*) alrededor del número 5, (*c*) alrededor de cero.

5.117. Encuentre los primeros cuatro momentos alrededor de la media de los números 0, 0, 0, 1, 1, 1, 1, 1.

5.118. (*a*) Demuestre que $m_5 = m_5' - 5m_1'm_4' + 10m_1'^2m_3' - 10m_1'^3m_2' + 4m_1'^5$. (*b*) Derive una fórmula similar para m_6.

5.119. De un total de n números, la fracción p está compuesta de unos, mientras que la fracción $q = 1 - p$ está compuesta de ceros. Encuentre (*a*) m_1, (*b*) m_2, (*c*) m_3, (*d*) m_4 para el conjunto de números. Compare con el problema 5.117.

5.120. Calcule los primeros cuatro momentos alrededor de la media para la distribución de la tabla 5-22.

Tabla 5-22

x	f
12	1
14	4
16	6
18	10
20	7
22	2
TOTAL	30

5.121. Calcule los primeros cuatro momentos alrededor de la media para la distribución del problema 5.97.

5.122. Encuentre (*a*) m_1, (*b*) m_2, (*c*) m_3, (*d*) m_4, (*e*) $\bar{x}$, (*f*) s, (*g*) $\overline{x^2}$, (*h*) $\overline{x^3}$, (*i*) $\overline{x^4}$, (*j*) $\overline{(x+1)^3}$ para la distribución del problema 5.100.

5.123. Encuentre los coeficientes de (*a*) sesgo, (*b*) curtosis, para la distribución del problema 5.120.

5.124. Encuentre los coeficientes de (*a*) sesgo (*b*) curtosis, para la distribución del problema 5.97. *Véase* el problema 5.121.

5.125. Los segundos momentos alrededor de la media de dos distribuciones son 9 y 16, mientras que los terceros momentos alrededor de la media son -8.1 y -12.8, respectivamente. ¿Cuál distribución tiene más sesgo hacia la izquierda?

5.126. Los cuartos momentos alrededor de la media de las dos distribuciones del problema 5.125 son 230 y 780, respectivamente. ¿Cuál distribución se aproxima más a la distribución normal desde el punto de vista de (*a*) su pico, (*b*) su sesgo?

PROBLEMAS DIVERSOS

5.127. Una población de 7 números tiene media 40 y desviación estándar 3. Si se toman muestras de tamaño 5 de esta población y se calcula la varianza de cada muestra, encuentre la media de la distribución muestral de varianzas si el muestreo se lleva a cabo (*a*) con remplazo, (*b*) sin remplazo.

5.128. Ciertos tubos que produce una fábrica tienen tiempo de vida media de 900 horas y desviación estándar de 80 horas. La fábrica despacha 1000 lotes de 100 tubos cada uno. ¿En cuántos lotes podemos esperar (*a*) que la vida media exceda las 910 horas? (*b*) que la desviación estándar del tiempo de vida excedan 95 horas?¿Qué supuestos se deben hacer?

5.129. En el problema 5.128 si la vida mediana es 900 horas, ¿en cuántos lotes podemos esperar que el tiempo de vida mediano exceda 910 horas? Compare su respuesta con el problema 5.128(*a*) y explique los resultados.

5.130. En una evaluación general las notas tuvieron distribución normal con media 72 y desviación estándar 8. (*a*) Encuentre la nota mínima del 20% superior de los estudiantes, (*b*) Encuentre la probabilidad de que en una muestra de 100 estudiantes tomada al azar, la nota mínima del 20% superior de los estudiantes sea menor que 76.

5.131. (*a*) Demuestre que la varianza del conjunto de n números $a, a + d, a + 2d, \ldots, a + (n - 1)d$ (es decir, la progresión aritmética con primer término a y diferencia común d) está dada por $\frac{1}{12}(n^2 - 1)d^2$. [*Ayuda*: Use $1 + 2 + 3 + \cdots + (n - 1) = \frac{1}{2}n(n - 1)$, $1^2 + 2^2 + 3^2 + \cdots + (n - 1)^2 = \frac{1}{6}n(n - 1)(2n - 1)$.] (*b*) Use (*a*) en el problema 5.103.

5.132. Demuestre que los primeros cuatro momentos alrededor de la media de la progresión aritmética $a, a + d, a + 2d, \ldots, a + (n - 1)d$ son

$$m_1 = 0, \qquad m_2 = \frac{1}{12}(n^2 - 1)d^2, \qquad m_3 = 0, \qquad m_4 = \frac{1}{240}(n^2 - 1)(3n^2 - 7)d^4$$

Compare con el problema 5.114. [*Ayuda*: $1^4 + 2^4 + 3^4 + \cdots + (n - 1)^4 = \frac{1}{30}n(n - 1)(2n - 1)(3n^2 - 3n - 1)$.]

Respuestas a los problemas suplementarios

5.49. (*a*) 9.0 (*b*) 4.47 (*c*) 9.0 (*d*) 3.16

5.50. (*a*) 9.0 (*b*) 4.47 (*c*) 9.0 (*d*) 2.58

5.51. (*a*) $\mu_{\bar{X}} = 22.40$ oz, $\sigma_{\bar{X}} = 0.008$ oz (*b*) $\mu_{\bar{X}} = 22.40$ oz, $\sigma_{\bar{X}}$ es ligeramente menor que 0.008 oz

5.52. (*a*) $\mu_{\bar{X}} = 22.40$ oz, $\sigma_{\bar{X}} = 0.008$ oz (*b*) $\mu_{\bar{X}} = 22.40$ oz, $\sigma_{\bar{X}} = 0.0057$ oz

5.53. (*a*) 237 (*b*) 2 (*c*) ninguno (*d*) 24

5.54. (*a*) 0.4972 (*b*) 0.1587 (*c*) 0.0918 (*d*) 0.9544

5.55. (*a*) 0.8164 (*b*) 0.0228 (*c*) 0.0038 (*d*) 1.0000

5.56. 0.0026

5.57. (*a*) 0.0029 (*b*) 0.9596 (*c*) 0.1446

5.58. (*a*) 2 (*b*) 996 (*c*) 218

5.59. (*a*) 0.0179 (*b*) 0.8664 (*c*) 0.1841

5.60. (*a*) 6 (*b*) 9 (*c*) 2 (*d*) 12

5.62. (*a*) 19 (*b*) 125

5.63. (*a*) 0.0077 (*b*) 0.8869

5.64. (*a*) 0.0028 (*b*) 0.9172

5.65. (*a*) 0.2150 (*b*) 0.0064 (*c*) 0.4504

5.66. 0.0.0482

5.67. 0.0188

5.68. 0.0410

5.70. (*a*) 118.79 lb (*b*) 0.74 lb

5.71. 0.0228

5.72. (*a*) 10.00 (*b*) 11.49

5.73. (*a*) 40/3 (*b*) 28.10

5.74. (*a*) 0.50 (*b*) 0.17 (*c*) 0.28

5.75. (*a*) 0.36 (*b*) 0.49

5.80. (*a*) entre 0.01 y 0.05 (*b*) mayor que 0.05

5.81. (*a*) mayor que 0.05 (*b*) mayor que 0.05

5.82. (*a*) 799 (*c*) 949.5 (*e*) 100 (horas) (*g*) 62/400 = 0.155 o 15.5% (*i*) 19.0%
(*b*) 1000 (*d*) 1099.5, 1199.5 (*f*) 76 (*h*) 29.5% (*j*) 78.0%

5.85. (*a*) 24% (*b*) 11% (*c*) 46%

5.86. (*a*) 0.003 centímetros
(*b*) 0.3195, 0.3225, 0.3255, . . . , 0.33 centímetros
(*c*) 0.320–0.322, 0.323–0.325, 0.326–0.328, . . . , 0.335–0.337

5.91. 86

5.92. 0.50 s

5.93. 8.25

5.94. (*a*) 82 (*b*) 79

5.95. 78

5.96. 80%, 20%

5.97. 11.09 ton

5.98. 501.0

5.99. 0.72642 centímetros

5.100. 26.2

5.101. (*a*) 2.16 (*b*) 0.90 (*c*) 0.484

5.103. 45

5.104. (*a*) 0.733 ton (*b*) 38.60

5.105. (*a*) $\bar{x} = 2.47$ (*b*) $s = 1.11$

5.106. (*a*) 0.000576 centímetros (*b*) 72.1%, 93.3%, 99.76%

5.107. (*a*) 146.8 lb, 12.9 lb

5.108. (*a*) 0.7349 centímetros, 0.00495 centímetros

5.110. (*a*) 6 (*b*) 40 (*c*) 288 (*d*) 2188

5.111. (*a*) 0 (*b*) 4 (*c*) 0 (*d*) 25.86

5.112. (*a*) −1 (*b*) 5 (*c*) −91 (*d*) 53

5.114. 0, 26.25, 0, 1193.1

5.115. 7

5.116. (*a*) 0, 6, 19, 42 (*b*) −4, 22, −117, 560 (*c*) 1, 7, 38, 155

5.117. 0, 0.2344, −0.0586, 0.0696

5.120. $m_1 = 0$, $m_2 = 5.97$, $m_3 = -3.97$, $m_4 = 89.22$

5.121. $m_1 = 0$, $m_2 = 0.53743$, $m_3 = 0.36206$, $m_4 = 0.84914$

5.122. (*a*) 0 (*c*) 92.35 (*e*) 26.2 (*g*) 739.38 (*i*) 706,428
(*b*) 52.95 (*d*) 7158.20 (*f*) 7.28 (*h*) 22,247 (*j*) 24,545

5.123. (*a*) −0.2464 (*b*) 2.62

5.124. (*a*) 0.9190 (*b*) 2.94

5.125. Primera distribución

5.126. (*a*) segunda (*b*) primera

5.127. (*a*) 7.2 (*b*) 8.4

5.128. (*a*) 106 (*b*) 4

5.129. 159

5.130. (*a*) 78.7 (*b*) 0.0090

CAPÍTULO 6

Teoría de la estimación

ESTIMACIÓN INSESGADA Y ESTIMACIÓN EFICIENTE

Como anotamos en el capítulo 5 (*véase* página 170), un estadístico recibe el nombre de *estimador insesgado* de un parámetro de una población si la media o la esperanza del estadístico es igual al parámetro. El valor correspondiente del estadístico se denomina, entonces, *estimador insesgado* del parámetro.

EJEMPLO 6.1. La media $\bar{X}$ y varianza $\hat{S}^2$, tal como se definieron en las páginas 167 y 170, son estimadores insesgados de la media de la población μ y varianza σ^2, ya que $E(\bar{X}) = \mu$, $E(\hat{S}^2) = \sigma^2$. Los valores $\bar{x}$ y $\hat{s}^2$ se denominan estimadores insesgados. Sin embargo, $\hat{S}$ es un estimador sesgado de σ, ya que, en general, $E(\hat{S}) \neq \sigma$.

Si las distribuciones muestrales de dos estadísticos tienen la misma media, el estadístico con la varianza menor se llama *estimador más eficiente* de la media. El valor correspondiente del estadístico eficiente recibe entonces el nombre de *estimador eficiente*. Es claro que en la práctica se prefiera tener estimadores que sean tanto eficientes como insesgados, pero esto no siempre es posible.

EJEMPLO 6.2. Para una población normal, la distribución muestral de la media y la mediana tienen la misma media; es decir, la media de la población. Sin embargo, la varianza de la distribución muestral de medias es más pequeña que aquélla de la distribución muestral de medianas. Por consiguiente, la media provee un estimador más eficiente que la mediana (*véase* tabla 5-1, página 172).

ESTIMACIÓN PUNTUAL Y ESTIMACIÓN POR INTERVALO. CONFIABILIDAD

Un estimador de un parámetro poblacional dado por un número sencillo se llama *estimador puntual* del parámetro. Un estimador de un parámetro poblacional dado por dos números entre los cuales se puede considerar que está el parámetro, se llama *estimación por intervalo* del parámetro.

EJEMPLO 6.3. Si decimos que una distancia es de 5.28 metros, estamos dando un estimador puntual. Si, por otro lado, decimos que la distancia es 5.28 ± 0.03 metros, es decir, la distancia se encuentra entre 5.25 y 5.31 metros, estamos dando una estimación por intervalo.

La *confiabilidad* de un estimador es el conocimiento de su error o de su precisión.

ESTIMACIÓN POR INTERVALO DE CONFIANZA DE PARÁMETROS DE LA POBLACIÓN

Sean μ_S y σ_S la media y la desviación estándar (error estándar) de la distribución muestral del estadístico S. Entonces, si la distribución muestral de S es aproximadamente normal (lo cual, como hemos visto, es verdadero para muchos estadísticos si el tamaño de la muestra $n \geqq 30$), podemos esperar que S se encuentre en el intervalo $\mu_S - \sigma_S$ a $\mu_S + \sigma_S$, $\mu_S - 2\sigma_S$ a $\mu_S + 2\sigma_S$ o $\mu_S - 3\sigma_S$ a $\mu_S + 3\sigma_S$ alrededor de 68.27%, 95.45% y 99.73% de las veces, respectivamente.

Asimismo podemos esperar encontrar, o podemos estar *confiados* de encontrar μ_S en los intervalos $S - \sigma_S$ a $S + \sigma_S$, $S - 2\sigma_S$ a $S + 2\sigma_S$ o $S - 3\sigma_S$ a $S + 3\sigma_S$ cerca del 68.27%, 95.45% y 99.73% de las veces, respectivamente. Debido a esto, llamamos a estos intervalos respectivos el 68.27%, 95.45% y 99.73% *intervalo de confianza* para el estimador de μ_S (es decir, para estimar el parámetro de la población en el caso de que S sea insesgada). Los números del extremo de estos intervalos ($S \pm \sigma_S$, $S \pm 2\sigma_S$, $S \pm 3\sigma_S$) se llaman entonces los *límites de confianza* del 68.27%, 95.45% y 99.73%.

De igual manera, $S \pm 1.96\sigma_S$ y $S \pm 2.58\sigma_S$ son los límites de confianza del 95% y 99% (o 0.95 y 0.99) de μ_S. El porcentaje de confianza recibe el nombre de *nivel de confianza*. Los números 1.96, 2.58, etc. en los límites de confianza se llaman *valores críticos* y se denotan como z_c. Podemos encontrar los valores críticos a partir de los niveles de confianza y viceversa.

En la tabla 6-1 damos los valores de z_c correspondientes a varios niveles de confianza que se utilizan en la práctica. Se pueden encontrar los valores de z_c para los niveles de confianza que no están en la tabla a partir del área de la curva normal en el apéndice C.

Tabla 6-1

Nivel de confianza	99.73%	99%	98%	96%	95.45%	95%	90%	80%	68.27%	50%
z_c	3.00	2.58	2.33	2.05	2.00	1.96	1.645	1.28	1.00	0.6745

En los casos en que un estadístico tiene una distribución muestral diferente de la distribución normal (tal como chi cuadrado, t o F), se deben hacer las modificaciones adecuadas para obtener los intervalos de confianza.

INTERVALO DE CONFIANZA PARA MEDIAS

1. **MUESTRAS GRANDES** ($n \geqq 30$). Si el estadístico S es la media muestral $\bar{X}$, entonces los límites de confianza del 95% y 99% para la estimación de la media de la población μ están dados por $\bar{X} \pm 1.96\sigma_{\bar{X}}$ y $\bar{X} \pm 2.58\sigma_{\bar{X}}$ respectivamente. De manera más general, los límites de confianza están dados por $\bar{X} \pm z_c\sigma_{\bar{X}}$ donde z_c, el cual depende del nivel particular deseado de confianza, puede obtenerse de la tabla anterior. Usando los valores de $\sigma_{\bar{X}}$ obtenidos en el capítulo 5, vemos que los límites de confianza de la media de la población están dados por

$$\bar{X} \pm z_c \frac{\sigma}{\sqrt{n}} \tag{1}$$

en el caso de que la toma de muestras sea de una población infinita o si el muestreo se lleva cabo con remplazo de una población finita, y por

$$\bar{X} \pm z_c \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \tag{2}$$

si el muestreo es sin remplazo de una población finita de tamaño N.

En general, la desviación estándar σ de la población es desconocida, de manera que para obtener los límites de confianza anteriores, debemos utilizar el estimador $\hat{S}$ o S.

2. MUESTRAS PEQUEÑAS ($n < 30$) **Y POBLACIÓN NORMAL.** En este caso usamos la distribución t para obtener los niveles de confianza. Por ejemplo, si $-t_{0.975}$ y $t_{0.975}$ son los valores de t para los cuales 2.5% del área se encuentra en cada cola de la distribución t, entonces un intervalo de confianza del 95% para T está dado por (*véase* página 171).

$$-t_{0.975} < \frac{(\bar{X} - \mu)\sqrt{n}}{\hat{S}} < t_{0.975} \tag{3}$$

a partir de lo cual vemos que μ puede calcularse de manera que quede en el intervalo

$$\bar{X} - t_{0.975}\frac{\hat{S}}{\sqrt{n}} < \mu < \bar{X} + t_{0.975}\frac{\hat{S}}{\sqrt{n}} \tag{4}$$

con 95% de confianza. En general, los límites de confianza para las medias de las poblaciones están dados por

$$\bar{X} \pm t_c\frac{\hat{S}}{\sqrt{n}} \tag{5}$$

donde el valor t_c se puede obtener del apéndice D.

Una comparación de (*5*) con (*1*) indica que para muestras pequeñas remplazamos z_c por t_c. Para $n \geqq 30$, z_c y t_c son prácticamente iguales. Se debe notar que una ventaja de la teoría de las muestras pequeñas (la cual, por supuesto, puede también usarse para muestras grandes, a saber, es *exacta*) es que $\hat{S}$ aparece en (*5*) de manea que la desviación estándar muestral puede usarse en lugar de la desviación estándar de la población (la que, por lo general, se desconoce) tal como en (*1*).

INTERVALO DE CONFIANZA PARA PROPORCIONES

Supongamos que el estadístico S es la proporción de "éxitos" en una muestra de tamaño $n \geqq 30$, obtenida a partir de una población con distribución binomial en la que p es la proporción de éxitos (es decir, la probabilidad de éxito). Entonces, los límites de confianza para p están dados por $P \pm z_c\sigma_P$, donde P es la proporción de éxitos en la muestra de tamaño n. Usando los valores de σ_p obtenidos en el capítulo 5, vemos que los límites de confianza para proporción de la población están dados por

$$P \pm z_c\sqrt{\frac{pq}{n}} = P \pm z_c\sqrt{\frac{p(1-p)}{n}} \tag{6}$$

en el caso en que el muestreo se haga de una población infinita o si se hace de una población finita y con reemplazo. De manera similar, los límites de confianza son

$$P \pm z_c\sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}} \tag{7}$$

si la toma de muestras se hace sin remplazo de una población de tamaño finito N. Observe que estos resultados se obtuvieron de (*1*) y (*2*) remplazando $\bar{X}$ por P y σ por $\sqrt{pq}$.

Para calcular los límites de confianza anteriores, usamos el estimado muestral P para p. En el problema 6.27 se da un método más exacto.

INTERVALO DE CONFIANZA PARA DIFERENCIAS Y SUMAS

Si S_1 y S_2 son dos estadísticos muestrales cuya distribución muestral es aproximadamente normal, los límites de confianza para las diferencias de los parámetros poblacionales correspondientes a S_1 y S_2 están dados por

$$S_1 - S_2 \pm z_c\sigma_{S_1 - S_2} = S_1 - S_2 \pm z_c\sqrt{\sigma_{S_1}^2 + \sigma_{S_2}^2} \tag{8}$$

mientras que los límites de confianza para la suma de los parámetros de la población están dados por

$$S_1 + S_2 \pm z_c\sigma_{S_1 + S_2} = S_1 + S_2 \pm z_c\sqrt{\sigma_{S_1}^2 + \sigma_{S_2}^2} \tag{9}$$

siempre y cuando las muestras sean independientes.

Por ejemplo, los límites de confianza para la diferencia de dos medias poblacionales, en el caso en que las poblaciones sean infinitas y tengan desviaciones estándar conocidas σ_1 y σ_2, están dados por

$$\bar{X}_1 - \bar{X}_2 \pm z_c \sigma_{\bar{X}_1 - \bar{X}_2} = \bar{X}_1 - \bar{X}_2 \pm z_c \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{10}$$

donde $\bar{X}_1, n_1$ y $\bar{X}_2, n_2$ son las medias y los tamaños respectivos de las dos muestras obtenidas de las poblaciones.

De manera similar, los límites de confianza para la diferencia de dos proporciones de poblaciones, donde las poblaciones son infinitas, están dados por

$$P_1 - P_2 \pm z_c \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}} \tag{11}$$

donde P_1 y P_2 son las dos proporciones muestrales y n_1 y n_2 son los tamaños de las dos muestras obtenidas de la población.

INTERVALO DE CONFIANZA PARA LA VARIANZA DE UNA DISTRIBUCIÓN NORMAL

El hecho de que $nS^2/\sigma^2 = (n-1)\hat{S}^2/\sigma^2$ tenga distribución chi cuadrado con $n-1$ grados de libertad, nos permite obtener límites de confianza para σ^2 o σ. Por ejemplo, si $\chi^2_{0.025}$ y $\chi^2_{0.975}$ son los valores de χ^2 para los cuales 2.5% del área está en cada cola de la distribución, entonces el 95% del intervalo de confianza es

$$\chi^2_{0.025} \leqq \frac{nS^2}{\sigma^2} \leqq \chi^2_{0.975} \tag{12}$$

o de manera equivalente

$$\chi^2_{0.025} \leqq \frac{(n-1)\hat{S}^2}{\sigma^2} \leqq \chi^2_{0.975} \tag{13}$$

De esto vemos que σ puede calcularse en el intervalo

$$\frac{S\sqrt{n}}{\chi_{0.975}} \leqq \sigma \leqq \frac{S\sqrt{n}}{\chi_{0.025}} \tag{14}$$

o de manera equivalente

$$\frac{\hat{S}\sqrt{n-1}}{\chi_{0.975}} \leqq \sigma \leqq \frac{\hat{S}\sqrt{n-1}}{\chi_{0.025}} \tag{15}$$

con el 95% de confianza. De igual manera, se pueden encontrar otros intervalos de confianza.

En general es deseable que la amplitud esperada de un intervalo de confianza sea lo más pequeña posible. Para estadísticos con distribuciones muestrales simétricas, tales como la normal y la t, esto se logra usando colas con áreas iguales. Sin embargo, para distribuciones no simétricas, tales como la chi cuadrado, puede ser deseable ajustar las áreas de las colas de manera que se obtenga el menor intervalo. El proceso se ilustra en el problema 6.28.

INTERVALO DE CONFIANZA PARA EL COCIENTE ENTRE VARIANZAS

En el capítulo 5, página 172, vimos que si dos variables aleatorias independientes de tamaños m y n con varianzas S_1^2, S_2^2 se obtienen de dos poblaciones con distribución normal de varianzas σ_1^2, σ_2^2, respectivamente, entonces la variable aleatoria $\dfrac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2}$ tiene distribución F con $m-1$, $n-1$ grados de libertad. Por ejemplo, si denotamos por $F_{0.01}$ y $F_{0.99}$ los valores de F para los cuales el 1% del área se encuentra en cada cola de la distribución F, entonces con el 98% de confianza tenemos

$$F_{0.01} \leqq \frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2} \leqq F_{0.99} \tag{16}$$

A partir de esto podemos ver que un intervalo de confianza del 98% para el cociente de varianzas σ_1^2/σ_2^2 de las dos poblaciones está dado por

$$\frac{1}{F_{0.99}}\frac{\hat{S}_1^2}{\hat{S}_2^2} \leqq \frac{\sigma_1^2}{\sigma_2^2} \leqq \frac{1}{F_{0.01}}\frac{\hat{S}_1^2}{\hat{S}_2^2} \tag{17}$$

Observe que $F_{0.99}$ se obtiene de una de las tablas del apéndice F. El valor $F_{0.01}$ es el recíproco de $F_{0.99}$ con los grados del numerador y el denominador invertidos, de acuerdo con el teorema 4-8, página 126.

De manera similar podemos encontrar un intervalo de confianza del 90% al usar la tabla apropiada del apéndice F. Esto estaría dado por

$$\frac{1}{F_{0.95}}\frac{\hat{S}_1^2}{\hat{S}_2^2} \leqq \frac{\sigma_1^2}{\sigma_2^2} \leqq \frac{1}{F_{0.05}}\frac{\hat{S}_1^2}{\hat{S}_2^2} \tag{18}$$

ESTIMACIÓN DE MÁXIMA VEROSIMILITUD

A pesar de que los límites de confianza son valiosos para estimar un parámetro de una población, conviene tener una estimación puntual o sencilla. Para obtener el "mejor" de tales estimadores, empleamos una técnica conocida como el *método de máxima verosimilitud*, debida a Fisher.

Para ilustrar el método, suponemos que la población tiene una función de densidad que contiene un parámetro poblacional, digamos θ, que debe estimarse por medio de algún estadístico. Entonces, la función de densidad puede denominarse $f(x, \theta)$. Suponiendo que hay n observaciones independientes. $X_1, \ldots, X_n$, la función de densidad conjunta para estas observaciones es

$$L = f(x_1, \theta)f(x_2, \theta)\cdots f(x_n, \theta) \tag{19}$$

lo cual se denomina *verosimilitud*. La *máxima verosimilitud* se puede obtener entonces tomando la derivada de L con respecto a θ e igualándola a cero. Para este propósito, es conveniente primero tomar los logaritmos y luego la derivada. Así encontramos

$$\frac{1}{f(x_1, \theta)}\frac{\partial f(x_1, \theta)}{\partial \theta} + \cdots + \frac{1}{f(x_n, \theta)}\frac{\partial f(x_n, \theta)}{\partial \theta} = 0 \tag{20}$$

La solución de esta ecuación, para θ en términos de x_k, se conoce como el *estimador de máxima verosimilitud de* θ.

El método es susceptible de generalización. En el caso de que haya varios parámetros, tomamos derivadas parciales con respecto a cada parámetro, las igualamos a cero y resolvemos las ecuaciones resultantes simultáneamente.

Problemas resueltos

ESTIMACIÓN INSESGADA Y ESTIMACIÓN EFICIENTE

6.1. Dé ejemplos de estimadores (o estimados) que sean (*a*) insesgados y eficientes, (*b*) insesgados e ineficientes, (*c*) sesgados e ineficientes.

Suponiendo que la población es normal. Entonces,

(*a*) La media muestral $\bar{X}$ y la varianza muestral modificada $\hat{S}^2 = \dfrac{n}{n-1}S^2$ son dos de tales ejemplos.

(*b*) La mediana muestral y el estadístico muestral $(\frac{1}{2})(Q_1 + Q_3)$, siendo Q_1 y Q_3 los cuartiles muestrales inferior y superior, son dos de tales ejemplos. Ambos estadísticos son estimadores insesgados de la media de la población, ya que se puede demostrar que la media de sus distribuciones muestrales es la media de la población. Sin embargo, ambos son ineficientes comparados con $\bar{X}$.

(*c*) Para estimar la desviación estándar de la población, σ, se pueden utilizar la desviación muestral estándar S, la desviación estándar modificada $\hat{S}$, la desviación media y el rango semiintercuartílico.

6.2. Un científico registró una muestra de cinco medidas del diámetro de una esfera como 6.33, 6.37, 6.36, 6.32 y 6.37 cm. Determine los estimadores insesgado y eficiente de (*a*) la verdadera media, (*b*) la verdadera varianza. Suponga que el diámetro medido tiene distribución normal.

(*a*) Un estimador insesgado y eficiente de la verdadera media (es decir, la media poblacional) es

$$\bar{x} = \frac{\sum x}{n} = \frac{6.33 + 6.37 + 6.36 + 6.32 + 6.37}{5} = 6.35\,\text{cm}$$

(*b*) Un estimador insesgado y eficiente de la verdadera varianza (es decir, la varianza poblacional) es

$$\hat{s}^2 = \frac{n}{n-1}\,s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

$$= \frac{(6.33-6.35)^2 + (6.37-6.35)^2 + (6.36-6.35)^2 + (6.32-6.35)^2 + (6.37-6.35)^2}{5-1}$$

$$= 0.00055\,\text{cm}^2$$

Observe que $\hat{s} = \sqrt{0.00055} = 0.023$ es un estimador de la verdadera desviación estándar, pero este estimador no es insesgado ni eficiente.

6.3. Del ejemplo de la página 173, suponga que la estatura de 100 estudiantes varones de la Universidad XYZ representa una muestra aleatoria de las estaturas de todos los 1546 estudiantes varones de dicha universidad. Determine los estimadores insesgados y eficientes de (*a*) la verdadera media, (*b*) la verdadera varianza.

(*a*) Del problema 5.33:

Estimador insesgado y eficiente de la verdadera estatura media $= \bar{x} = 67.45$ pulgadas

(*b*) Del problema 5.38:

Estimador insesgado y eficiente de la verdadera varianza $= \hat{s}^2 = \frac{n}{n-1}\,s^2 = \frac{100}{99}\,(8.5275) = 8.6136$

Por consiguiente, $\hat{s} = \sqrt{8.6136} = 2.93$. Observe que cuando n es grande no hay diferencia entre s^2 y $\hat{s}^2$ o entre s y $\hat{s}$.

6.4. Dé un estimado insesgado e ineficiente del verdadero diámetro (media) de la esfera del problema 6.2.

La mediana es un ejemplo de un estimador insesgado e ineficiente de la media de la población. Para las cinco medidas organizadas en orden de magnitud, la mediana es 6.36 cm.

ESTIMACIÓN DEL INTERVALO DE CONFIANZA PARA MEDIAS (MUESTRAS GRANDES)

6.5. Encuentre los intervalos de confianza del (*a*) 95%, (*b*) 99% para estimar la estatura media de los estudiantes de la Universidad XYZ del problema 6.3.

(*a*) Los límites de confianza del 95% son $\bar{X} \pm 1.96\sigma/\sqrt{n}$.

Usando $\bar{x} = 67.45$ pulgadas y $\hat{s} = 2.93$ pulgadas como estimado de σ (*véase* el problema 6.3), los límites de confianza son $67.45 \pm 1.96(2.93/\sqrt{100})$ o 67.45 ± 0.57 pulgadas. Entonces, el intervalo de confianza del 95% para la media de la población μ es de 66.88 a 68.02 pulgadas, lo cual puede escribirse así: $66.88 < \mu < 68.02$.

Por consiguiente, podemos decir que la probabilidad de que la media de la estatura de la población se encuentre entre 66.88 y 68.02 pulgadas es de alrededor del 95% o 0.95. En símbolos escribimos $P(66.88 < \mu < 68.02) = 0.95$. Esto equivale a decir que tenemos 95% de *confianza* de que la media de la población (o verdadera media) se encuentre entre 66.88 y 68.02 pulgadas.

(*b*) Los límites de confianza del 99% son $\bar{X} \pm 2.58\sigma/\sqrt{n}$. Para el ejemplo dado,

$$\bar{x} \pm 2.58\,\frac{\hat{s}}{\sqrt{n}} = 67.45 \pm 2.58\,\frac{2.93}{\sqrt{100}} = 67.45 \pm 0.76\,\text{pulgadas}$$

Por consiguiente, el intervalo de confianza del 99% para la media poblacional μ es de 66.69 a 68.21 pulgadas, lo cual puede escribirse: $66.69 < \mu < 68.21$.

Para obtener los intervalos de confianza anteriores suponemos que la población es infinita o tan grande que podemos considerar que las condiciones son las mismas que cuando se toman muestras con remplazo. Para poblaciones finitas donde el muestreo se lleva a cabo sin remplazo, debemos usar $\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ en lugar de $\frac{\sigma}{\sqrt{n}}$. Sin embargo, podemos considerar el factor $\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{1546-100}{1546-1}} = 0.967$ como 1.0, de manera que no sea necesario usarlo. Si se usa, los límites de confianza anteriores se convierten en 67.45 $\pm$0.56 y 67.45 $\pm$0.73 pulgadas, respectivamente.

6.6. Las medidas de los diámetros de una muestra aleatoria de 200 balineras hechas por cierta máquina durante una semana mostraron una media de 0.824 centímetros y desviación estándar 0.042 centímetros. Encuentre los límites de confianza del (*a*) 95%, (*b*) 99% para el diámetro medio de todas las balineras.

Ya que $n = 200$ es grande, podemos suponer que $\bar{X}$ tiene distribución casi normal.

(*a*) Los límites de confianza del 95% son

$$\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}} = \bar{x} \pm 1.96\frac{\hat{s}}{\sqrt{n}} = 0.824 \pm 1.96\frac{0.042}{\sqrt{200}} = 0.824 \pm 0.0058 \text{ centímetros}$$

o 0.824 $\pm$0.006 centímetros.

(*b*) Los límites de confianza del 99% son

$$\bar{X} \pm 2.58\frac{\sigma}{\sqrt{n}} = \bar{x} \pm 2.58\frac{\hat{s}}{\sqrt{n}} = 0.824 \pm 2.58\frac{0.042}{\sqrt{200}} = 0.824 \pm 0.0077 \text{ centímetros}$$

o 0.824 $\pm$0.008 centímetros.

Observe que supusimos que la desviación estándar reportada es la desviación estándar modificada $\hat{s}$. Si la desviación estándar hubiera sido s, habríamos usado $\hat{s} = \sqrt{n/(n-1)}\,s = \sqrt{200/199}\,s$ lo cual puede tomarse como s para todos los propósitos prácticos. En general, para $n \geqq 30$, podemos tomar s y $\hat{s}$ como prácticamente iguales.

6.7. Encuentre los límites de confianza del (*a*) 98%, (*b*) 90%, (*c*) 99.73% del diámetro medio de las balineras del problema 6.6.

(*a*) Sea z_c tal que el área bajo la curva normal a la derecha de $z = z_c$ sea 1%. Entonces, por simetría, el área a la izquierda de $z = -z_c$ también es 1%, de manera que el área sombreada es el 98% del área total (figura 6-1).

Ya que el área total bajo la curva es uno, el área desde $z = 0$ es $z = z_c$ es 0.49; de ahí, $z_c = 2.33$.

Por consiguiente, los límites de confianza del 98% son

$$\bar{x} \pm 2.33\frac{\sigma}{\sqrt{n}} = 0.824 \pm 2.33\frac{0.042}{\sqrt{200}} = 0.824 \pm 0.0069 \text{ centímetros}$$

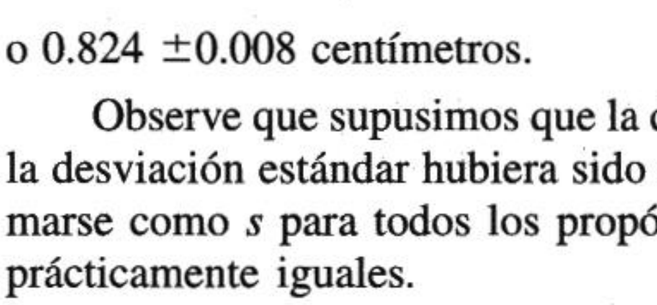


Fig. 6-1

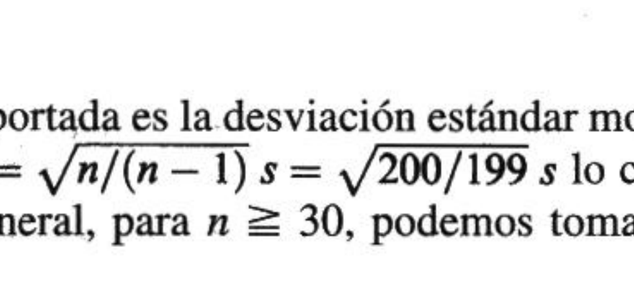


Fig. 6-2

(*b*) Necesitamos z_c tal que el área desde $z = 0$ a $z = z_c$ sea 0.45; entonces $z_c = 1.645$ (figura 6-2).

Por consiguiente, los límites de confianza del 90% son

$$\bar{x} \pm 1.645\frac{\sigma}{\sqrt{n}} = 0.824 \pm 1.645\frac{0.042}{\sqrt{200}} = 0.824 \pm 0.0049 \text{ centímetros}$$

(*c*) Los límites de confianza del 99.73% son

$$\bar{x} \pm 3\frac{\sigma}{\sqrt{n}} = 0.824 \pm 3\frac{0.042}{\sqrt{200}} = 0.824 \pm 0.0089 \text{ centímetros}$$

6.8. Al medir un tiempo de reacción, un psicólogo estima que la desviación estándar es 0.05 segundos. ¿Qué tan grande debe ser la muestra de medidas que él tome para que tenga la confianza del (*a*) 95%, (*b*) 99% de que el error en el estimador de la media del tiempo de reacción no exceda 0.01 segundos?

(*a*) Los límites de confianza del 95% son $\bar{X} \pm 1.96\sigma/\sqrt{n}$, siendo el error del estimado $1.96\sigma/\sqrt{n}$. Tomando $\sigma = s = 0.05$ segundos, vemos que este error será igual a 0.01 segundos si $(1.96)(0.05)/\sqrt{n} = 0.01$, es decir, $\sqrt{n} = (1.96)(0.05)/0.01 = 9.8$, o $n = 96.04$. Por consiguiente, podemos tener el 95% de confianza de que el error en el estimador será menor que 0.01 si $n = 97$ o más grande.

(*b*) Los límites de confianza del 99% son $\bar{X} \pm 2.58\sigma/\sqrt{n}$. Entonces $(2.58)(0.05)/\sqrt{n} = 0.01$, o $n = 166.4$. Por consiguiente, tenemos el 99% de confianza de que el error en el estimador será menor que 0.01 si y sólo si n es 167 o más grande.

Observe que las soluciones anteriores suponen una distribución casi normal para $\bar{X}$, lo cual se justifica ya que la n obtenida es grande.

6.9. Una muestra aleatoria de 50 notas de matemáticas de un total de 200 mostraron una media de 75 y desviación estándar de 10. (*a*) ¿Cuáles son los límites de confianza del 95% para la media de 200 notas? (*b*) ¿Con qué grado de confianza podemos decir que la media de las 200 notas es 75 ± 1?

(*a*) Dado que el tamaño de la población no es muy grande comparado con el tamaño de la muestra, debemos hacer los ajustes para el *muestreo sin remplazo*. Entonces, los límites de confianza del 95% son

$$\bar{X} \pm 1.96\sigma_{\bar{X}} = \bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = 75 \pm 1.96\frac{10}{\sqrt{50}}\sqrt{\frac{200-50}{200-1}} = 75 \pm 2.4$$

(*b*) Los límites de confianza pueden representarse por

$$\bar{X} \pm z_c\sigma_{\bar{X}} = \bar{X} \pm z_c\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = 75 \pm z_c\frac{(10)}{\sqrt{50}}\sqrt{\frac{200-50}{200-1}} = 75 \pm 1.23z_c$$

Ya que esto debe ser igual a 75 ± 1, tenemos $1.23\,z_c = 1$ o $z_c = 0.81$. El área bajo la curva normal desde $z = 0$ a $z = 0.81$ es 0.2910; de allí que el grado de confianza requerido sea $2(0.2919) = 0.582$ o 58.2%.

ESTIMACIÓN DEL INTERVALO DE CONFIANZA PARA MEDIAS (MUESTRAS PEQUEÑAS)

6.10. Los valores crítico del 95% (dos colas) para la distribución normal están dados por ± 1.96. ¿Cuáles son los valores correspondientes para la distribución t si el número de grados de libertad es (*a*) $v = 9$, (*b*) $v = 20$, (*c*) $v = 30$, (*d*) $v = 60$?

Para valores críticos del 95% (dos colas) el área sombreada total en la figura 6-3 debe ser 0.05. Por consiguiente, el área sombreada total en la cola derecha es 0.025 y el valor crítico correspondiente es $t_{0.975}$. Entonces, los valores críticos requeridos son $\pm\, t_{0.975}$. Para los valores dados de v éstos son (*a*) ± 2.26, (*b*) ± 2.09, (*c*) ± 2.04, (*d*) ± 2.00.

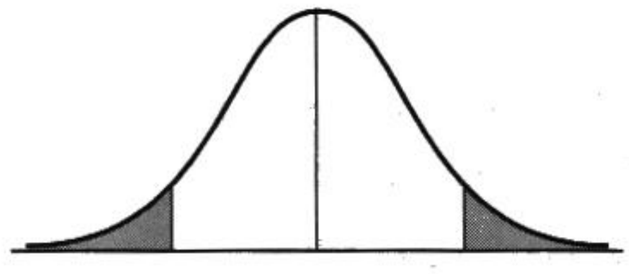


Fig. 6-3

6.11. Una muestra de 10 medidas del diámetro de una esfera dio una media $\bar{x} = 4.38$ centímetros y desviación estándar $s = 0.06$ centímetros. Encuentre los límites de confianza del (*a*) 95% y (*b*) 99% para el verdadero diámetro.

(*a*) Los límites del 95% están dados por $\bar{X} \pm t_{0.975}(S/\sqrt{n-1})$.

Ya que $v = n - 1 = 10 - 1 = 9$, encontramos que $t_{0.975} = 2.26$ [*véase también* el problema 6.10(*a*)]. Entonces, usando $\bar{x} = 4.38$ y $s = 0.06$, los límites de confianza del 95% requeridos son

$$4.38 \pm 2.26 \frac{0.06}{\sqrt{10-1}} = 4.38 \pm 0.0452 \text{ centímetros}$$

Por consiguiente, podemos tener el 95% de confianza de que las verdaderas medias estarán entre $4.38 - 0.045 = 4.335$ centímetros y $4.38 + 0.045 = 4.425$ centímetros.

(*b*) Para $v = 9$, $t_{0.99} = 3.25$. Entonces los límites de confianza del 99% son

$$\bar{X} \pm t_{0.995}(S/\sqrt{n-1}) = 4.38 \pm 3.25(0.06/\sqrt{10-1}) = 4.38 \pm 0.0650 \text{ centímetros}$$

y el intervalo de confianza del 99% es de 4.315 a 4.445 centímetros.

6.12. (*a*) Desarrolle el problema 6.11 suponiendo que son válidos los métodos de la teoría de las muestras grandes. (*b*) Compare los resultados de los dos métodos.

(*a*) Usando la teoría de las muestras grandes, los límites de confianza del 95% son

$$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}} = 4.38 \pm 1.96 \frac{0.06}{\sqrt{10}} = 4.38 \pm 0.037 \text{ centímetros}$$

donde usamos la desviación estándar muestral 0.06 como estimador de σ. De manera similar, los límites de confianza del 99% son $4.38 \pm (2.58)(0.06)/\sqrt{10} = 4.38 \pm 0.049$ centímetros.

(*b*) En cada caso los intervalos de confianza que utilizan métodos para muestras pequeñas o exactas son más amplios que aquéllos obtenidos usando métodos para muestras grandes. Esto es de esperarse ya que se obtiene menor precisión con muestras pequeñas que con muestras grandes.

ESTIMACIÓN DEL INTERVALO DE CONFIANZA PARA PROPORCIONES

6.13. Una encuesta de 100 votantes, escogidos al azar de todos los votantes en un distrito dado, indicó que el 55% está a favor de un candidato en particular. Encuentre los límites de confianza del (*a*) 95%, (*b*) 99%, (*c*) 99.73% de la proporción de todos los votantes a favor de ese candidato.

(*a*) Los límites de confianza del 95% de la p poblacional son

$$P \pm 1.96\sigma_P = P \pm 1.96\sqrt{\frac{p(1-p)}{n}} = 0.55 \pm 1.96\sqrt{\frac{(0.55)(0.45)}{100}} = 0.55 \pm 0.10$$

donde usamos la proporción muestral de 0.55 para estimar p.

(*b*) Los límites de confianza del 99% para p son $0.55 \pm 2.58\sqrt{(0.55)(0.45)/100} = 0.55 \pm 0.13$.

(*c*) Los límites de confianza del 99.73% para p son $0.55 \pm 3\sqrt{(0.55)(0.45)/100} = 0.55 \pm 0.15$.

Para desarrollar este problema con un método más exacto, *véase* problema 6.27.

6.14. ¿Qué tan grande debe ser la muestra de votantes que debe tomarse en el problema 6.13 para tener el 95% de confianza de que el candidato será electo?

El candidato es electo si $p > 0.50$, y para que tengamos confianza del 95% de su elección, necesitamos que Prob. $(p > 0.50) = 0.95$. Ya que $(P - p)/\sqrt{p(1-p)/n}$ es normal asintóticamente.

$$\text{Prob.}\left(\frac{P-p}{\sqrt{p(1-p)/n}} < \beta\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\beta} e^{-u^2/2}\,du$$

o

$$\text{Prob.}(p > P - \beta\sqrt{p(1-p)/n}) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\beta} e^{-u^2/2}\,du$$

La comparación con Prob. $(p > 0.50) = 0.95$, usando el apéndice C, muestra que

$$P - \beta\sqrt{p(1-p)/n} = 0.50 \qquad \text{donde} \qquad \beta = 1.645$$

Entonces, usando $P = 0.55$ y el estimador $p = 0.55$ del problema 6.13 tenemos

$$0.55 - 1.645\sqrt{(0.55)(0.45)/n} = 0.50 \qquad \text{o} \qquad n = 271$$

6.15. En 40 lanzamientos de una moneda se obtuvieron 24 caras. Encuentre los límites de confianza del (*a*) 95%, (*b*) 99.73% para la proporción de caras que se obtendrán en un número ilimitado de lanzamientos de la moneda.

(*a*) Al nivel del 95%, $z_c = 1.96$. Sustituyendo los valores $P = 24/40 = 0.6$ y $n = 40$ en la fórmula $p = P \pm z_c\sqrt{P(1-P)/n}$, encontramos $p = 0.60 \pm 0.15$, obteniéndose el intervalo de 0.45 a 0.75.

(*b*) Al nivel del 99.73%, $z_c = 3$. Usando la fórmula $p = P \pm z_c\sqrt{P(1-P)/n}$, encontramos $p = 0.60 \pm 0.23$, obteniéndose el intervalo de 0.37 a 0.83.

La fórmula más exacta del problema 6.27 da los intervalos de confianza del 95% como de 0.45 a 0.74 y los del 99.73% como de 0.37 a 0.79.

INTERVALO DE CONFIANZA PARA DIFERENCIAS Y SUMAS

6.16. Una muestra de 150 bombillas de la marca *A* mostró tener un tiempo de vida media de 1400 horas y desviación estándar de 120 horas. La muestra de 200 bombillas de la marca *B* mostró tener un tiempo de vida media de 1200 horas y desviación estándar de 80 horas. Encuentre los límites de confianza del (*a*) 95%, (*b*) 99% para la diferencia de los tiempos de vida media de las poblaciones de las marcas *A* y *B*.

Los límites de confianza para la diferencia en medias de las marcas *A* y *B* están dados por

$$\bar{X}_A - \bar{X}_B \pm z_c\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$$

(*a*) Los límites de confianza del 95% son $1400 - 1200 \pm 1.96\sqrt{(120)^2/150 + (80)^2/100} = 200 \pm 24.8$.

Por consiguiente, podemos tener confianza del 95% de que la diferencia de las medias de las poblaciones esté entre 175 y 225 horas.

(*b*) Los límites de confianza del 99% son $1400 - 1200 \pm 2.58\sqrt{(120)^2/150 + (80)^2/100} = 200 \pm 32.6$.

Por consiguiente, podemos tener confianza del 99% de que la diferencia de las medias de las poblaciones esté entre 167 y 233 horas.

6.17. En una muestra aleatoria de 400 adultos y 600 adolescentes que veían cierto programa de televisión, 100 adultos y 300 adolescentes dijeron que les gustó. Construya los límites de confianza del (*a*) 95%, (*b*) 99% para la diferencia en proporciones de todos los adultos y adolescentes que vieron el programa y lo encontraron de su agrado.

Los límites de confianza para la diferencia en proporciones de los dos grupos están dados por

$$P_1 - P_2 \pm z_c\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$$

Donde los subíndices 1 y 2 se refieren a adolescentes y adultos, respectivamente, y $Q_1 = 1 - P_1$, $Q_2 = 1 - P_2$ los complementos. Aquí $P_1 = 300/600 = 0.50$ y $P_2 = 100/400 = 0.25$ son, respectivamente, la proporción de adolescentes y adultos a los que les gustó el programa.

(*a*) Límites de confianza del 95%: $0.50 - 0.25 \pm 1.96\sqrt{(0.50)(0.50)/600 + (0.25)(0.75)/400} = 0.25 \pm 0.06$.

Por consiguiente, podemos tener confianza del 95% de que la diferencia verdadera de proporciones esté entre 0.19 y 0.31.

(*b*) Límites de confianza del 99%: $0.50 - 0.25 \pm 2.58\sqrt{(0.50)(0.50)/600 + (0.25)(0.75)/400} = 0.25 \pm 0.08$.

Por consiguiente, podemos tener confianza del 99% de que la diferencia verdadera de proporciones esté entre 0.17 y 0.33.

6.18. La fuerza electromotriz (fem) de las baterías producidas por una compañía tiene distribución normal con media 45.1 voltios y desviación estándar 0.04 voltios. Si cuatro de esas baterías se conectan en serie, encuentre los límites de confianza del (*a*) 95%, (*b*) 99%, (*c*) 99.73% de la fuerza electromotriz total.

Si E_1, E_2, E_3 y E_4 representan la fem de las cuatro baterías, tenemos

$$\mu_{E_1+E_2+E_3+E_4} = \mu_{E_1} + \mu_{E_2} + \mu_{E_3} + \mu_{E_4} \qquad \text{y} \qquad \sigma_{E_1+E_2+E_3+E_4} = \sqrt{\sigma_{E_1}^2 + \sigma_{E_2}^2 + \sigma_{E_3}^2 + \sigma_{E_4}^2}$$

Entonces, ya que $\mu_{E_1} = \mu_{E_2} = \mu_{E_3} = \mu_{E_4} = 45.1$ voltios y $\sigma_{E_1} = \sigma_{E_2} = \sigma_{E_3} = \sigma_{E_4} = 0.04$ voltios,

$$\mu_{E_1+E_2+E_3+E_4} = 4(45.1) = 180.4 \qquad \text{y} \qquad \sigma_{E_1+E_2+E_3+E_4} = \sqrt{4(0.04)^2} = 0.08$$

(*a*) Los límites de confianza del 95% son: $180.4 \pm 1.96(0.08) = 180.4 \pm 0.16$ voltios.

(*b*) Los límites de confianza del 99% son: $180.4 \pm 2.58(0.08) = 180.4 \pm 0.21$ voltios.

(*c*) Los límites de confianza del 99.73% son: $180.4 \pm 3(0.08) = 180.4 \pm 0.24$ voltios.

(*d*) Los límites de confianza del 50% son: $180.4 \pm 0.6745(0.08) = 180.4 \pm 0.054$ voltios.

El valor 0.054 voltios recibe el nombre de error probable.

INTERVALO DE CONFIANZA PARA VARIANZAS

6.19. Se calculó que la desviación estándar de los tiempos de vida de una muestra de 200 bombillas eléctricas es 100 horas. Encuentre los límites de confianza del (*a*) 95%, (*b*) 99% para la desviación estándar de todas esas bombillas eléctricas.

En este caso se aplica la teoría de muestreo grande. Por consiguiente (*véase* tabla 5-1, página 172) los límites de confianza para la desviación estándar de la población σ están dados por $S \pm z_c\sigma/\sqrt{2n}$, donde z_c indica el nivel de confianza. Usamos la desviación muestral estándar para calcular σ.

(*a*) Los límites de confianza del 95% son $100 \pm 1.96(100)/\sqrt{400} = 100 \pm 9.8$.

Por consiguiente, podemos tener confianza del 95% de que la desviación estándar de la población estará entre 90.2 y 109.8 horas.

(*b*) Los límites de confianza del 99% son $100 \pm 2.58(100)/\sqrt{400} = 100 \pm 12.9$.

Por consiguiente, podemos tener confianza del 99% de que la desviación estándar de la población estará entre 87.1 y 112.9 horas.

6.20. ¿Qué tan grande debe ser la muestra que tomemos de bombillas eléctricas del problema 6.19, para que tengamos confianza del 99.73% de que la desviación estándar de la población verdadera no difiere de la desviación estándar muestral por más de (*a*) 5%, (*b*) 10%?

Tal como en el problema 6.19, los límites de confianza del 99.73% para σ son $S \pm 3\sigma/\sqrt{2n} = s \pm 3s/\sqrt{2n}$, usando s como estimado de σ. Entonces el porcentaje de error en la desviación estándar es

$$\frac{3s/\sqrt{2n}}{s} = \frac{300}{\sqrt{2n}}\%$$

(*a*) Si $300/\sqrt{2n} = 5$, entonces $n = 1800$. Por consiguiente, el tamaño de la muestra debe ser de 1800 o más.

(*b*) Si $300/\sqrt{2n} = 10$, entonces $n = 450$. Por consiguiente, el tamaño de la muestra debe ser de 450 o más.

6.21. La desviación estándar de las estaturas de 16 estudiantes varones, seleccionados al azar en un colegio de 1000 estudiantes varones, es 2.40 pulgadas. Encuentre los límites de confianza del (*a*) 95%, (*b*) 99% de la desviación estándar de todos los estudiantes varones del colegio. Suponga que la estatura tiene distribución normal.

(*a*) Los límites de confianza del 95% están dados por $S\sqrt{n}/\chi_{0.975}$ y $S\sqrt{n}/\chi_{0.025}$.

Para $\nu = 16 - 1 = 15$ grados de libertad, $\chi^2_{0.975} = 27.5$ o $\chi_{0.975} = 5.24$ y $\chi^2_{0.025} = 6.26$ $\chi_{0.025} = 2.50$.

Entonces, los límites de confianza del 95% son $2.40\sqrt{16}/5.24$ y $2.40\sqrt{16}/2.50$, es decir, 1.83 y 3.84 pulgadas.

Por consiguiente, podemos tener confianza del 95% de que la desviación estándar de la población esté entre 1.83 y 3.84 pulgadas.

(*b*) Los límites de confianza del 99% están dados por $S\sqrt{n}/\chi_{0.995}$ y $S\sqrt{n}/\chi_{0.005}$.

Para $\nu = 16 - 1 = 15$ grados de libertad, $\chi^2_{0.995} = 32.8$ o $\chi_{0.995} = 5.73$ y $\chi^2_{0.005} = 4.60$ $\chi_{0.005} = 21.4$.

Entonces, los límites de confianza del 99% son $2.40\sqrt{16}/5.73$ y $2.40\sqrt{16}/2.14$, es decir, 1.68 y 4.49 pulgadas.

Por consiguiente, podemos tener confianza del 99% de que la desviación estándar de la población esté entre 1.68 y 4.49 pulgadas.

6.22. Desarrolle el problema 6.19 usando la teoría de muestras pequeñas o exactas.

(*a*) El límite de confianza del 95% está dado por $S\sqrt{n}/\chi_{0.975}$ y $S\sqrt{n}/\chi_{0.025}$.

Para $\nu = 200 - 1 = 199$ grados de libertad encontramos que, tal como en el problema 4.41, página 148,

$$\chi^2_{0.975} = \frac{1}{2}(z_{0.975} + \sqrt{2(199) - 1})^2 = \frac{1}{2}(1.96 + 19.92)^2 = 239$$

$$\chi^2_{0.025} = \frac{1}{2}(z_{0.025} + \sqrt{2(199) - 1})^2 = \frac{1}{2}(-1.96 + 19.92)^2 = 161$$

a partir de lo cual $\chi_{0.975} = 15.5$ y $\chi_{0.025} = 12.7$.

Entonces los límites de confianza del 95% son $100\sqrt{200}/15.5 = 91.2$ y $100\sqrt{200}/12.7 = 111.3$ horas respectivamente. Por consiguiente, podemos tener confianza del 95% de que la desviación estándar de la población estará entre 91.2 y 111.3 horas.

Esto debe compararse con el resultado del problema 6.19(*a*).

(*b*) El límite de confianza del 99% está dado por $S\sqrt{n}/\chi_{0.995}$ y $S\sqrt{n}/\chi_{0.005}$.

Para $\nu = 200 - 1 = 199$ grados de libertad,

$$\chi^2_{0.995} = \frac{1}{2}(z_{0.995} + \sqrt{2(199) - 1})^2 = \frac{1}{2}(2.58 + 19.92)^2 = 253$$

$$\chi^2_{0.005} = \frac{1}{2}(z_{0.005} + \sqrt{2(199) - 1})^2 = \frac{1}{2}(-2.58 + 19.92)^2 = 150$$

a partir de lo cual $\chi_{0.995} = 15.9$ y $\chi_{0.005} = 12.2$.

Entonces los límites de confianza del 99% son $100\sqrt{200}/15.9 = 88.9$ y $100\sqrt{200}/12.2 = 115.9$ horas, respectivamente. Por consiguiente, podemos tener confianza del 99% de que la desviación estándar de la población estará entre 88.9 y 115.9 horas.

Esto debe compararse con el resultado del problema 6.19(*b*).

INTERVALO DE CONFIANZA PARA EL COCIENTE ENTRE VARIANZAS

6.23. Se obtienen al azar dos muestras de tamaños 16 y 10, respectivamente, de dos poblaciones con distribución normal. Si sus varianzas son 24 y 18, respectivamente, encuentre los límites de confianza del (*a*) 98%, (*b*) 90% para el cociente de las varianzas.

(*a*) Tenemos $m = 16$, $n = 10$, $s_1^2 = 20$, $s_2^2 = 18$, así que

$$\hat{s}_1^2 = \frac{m}{m-1}s_1^2 = \left(\frac{16}{15}\right)(24) = 25.2$$

$$\hat{s}_2^2 = \frac{n}{n-1}s_2^2 = \left(\frac{10}{9}\right)(18) = 20.0$$

Del problema 4.47(*b*), página 151, tenemos $F_{0.99} = 4.96$ para $\nu_1 = 16 - 1 = 15$ y $\nu_2 = 10 - 1 = 9$ grados de libertad. También, del problema 4.47(*d*) tenemos para $\nu_1 = 15$ y $\nu_2 = 9$ grados de libertad que $F_{0.01} = 1/3.89$, así que $1/F_{0.01} = 3.89$. Entonces usando (*17*), página 213, encontramos que para el intervalo de confianza del 98%

$$\left(\frac{1}{4.96}\right)\left(\frac{25.2}{20.0}\right) \leqq \frac{\sigma_1^2}{\sigma_2^2} \leqq (3.89)\left(\frac{25.2}{20.0}\right)$$

o

$$0.283 \leqq \frac{\sigma_1^2}{\sigma_2^2} \leqq 4.90$$

(*b*) Como en (*a*), encontramos del apéndice F que $F_{0.95} = 3.01$ y $F_{0.05} = 1/2.59$. Por consiguiente, el intervalo de confianza del 90% es

$$\frac{1}{3.01}\left(\frac{25.2}{20.0}\right) \leqq \frac{\sigma_1^2}{\sigma_2^2} \leqq (2.59)\left(\frac{25.2}{20.0}\right)$$

o

$$0.4186 \leqq \frac{\sigma_1^2}{\sigma_2^2} \leqq 3.263$$

Observe que el intervalo de confianza del 90% es mucho más pequeño que el intervalo de confianza del 98%, tal como podríamos esperar, por supuesto.

6.24. Encuentre los límites de confianza de (*a*) 98%, (*b*) 90% para el cociente de las desviaciones estándar del problema 6.23.

Tomando las raíces cuadrado de las desigualdades del problema 6.23 encontramos los límites de confianza del 98% y 90%

(*a*) $$0.53 \leqq \frac{\sigma_1}{\sigma_2} \leqq 2.21$$

(*b*) $$0.65 \leqq \frac{\sigma_1}{\sigma_2} \leqq 1.81$$

ESTIMACIÓN DE MÁXIMA VEROSIMILITUD

6.25. Supongamos que se hacen n observaciones, $X_1, \ldots, X_n$ de una población con distribución normal, de la cual se desconoce la media mientras que la varianza es conocida. Encuentre el estimador de máxima verosimilitud de la media.

Ya que $$f(x_k, \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_k-\mu)^2/2\sigma^2}$$

tenemos

(*1*) $$L = f(x_1, \mu) \cdots f(x_n, \mu) = (2\pi\sigma^2)^{-n/2} e^{-\sum (x_k-\mu)^2/2\sigma^2}$$

Por consiguiente,

(*2*) $$\ln L = -\frac{n}{2} \ln (2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum (x_k - \mu)^2$$

Tomando la derivada parcial con respecto a μ se obtiene

(*3*) $$\frac{1}{L}\frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2} \sum (x_k - \mu)$$

remplazando $\partial L/\partial \mu = 0$ se obtiene

(*4*) $$\sum (x_k - \mu) = 0 \text{ ; es decir, } \sum x_k - n\mu = 0$$

o

$$\mu = \frac{\sum x_k}{n} \tag{5}$$

Por consiguiente, el estimador de máxima verosimilitud es la media muestral.

6.26. Si en el problema 6.25 se conoce la media pero la varianza es desconocida, encuentre el estimador de máxima verosimilitud de la varianza.

Si escribimos $f(x_k, \sigma^2)$ en lugar de $f(x_k, \mu)$, todo lo hecho en el problema 6.25 con las ecuaciones (2) aún aplica. Entonces, tomando la derivada parcial con respecto a σ^2, tenemos

$$\frac{1}{L}\frac{\partial L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum(x_k - \mu)^2$$

Remplazando $\partial L/\partial \sigma^2 = 0$, encontramos

$$\sigma^2 = \frac{\sum(x_k - \mu)^2}{n}$$

PROBLEMAS DIVERSOS

6.27. (*a*) Si P es la proporción observada de éxitos en una muestra de tamaño n, demuestre que los límites de confianza para calcular la proporción de éxitos de p, a un nivel de confianza determinado por z_c, está dado por

$$\frac{P + \frac{z_c^2}{2n} \pm z_c\sqrt{\frac{P(1-P)}{n} + \frac{z_c^2}{4n^2}}}{1 + \frac{z_c^2}{n}}$$

(*b*) Use la fórmula derivada en (*a*) para obtener los límites de confianza del 99.73% del problema 6.13. (*c*) Demuestre que para n grande la fórmula en (*a*) se reduce a $P \pm z_c\sqrt{P(1-P)/n}$, como se usó en el problema 6.13.

(*a*) La proporción de la muestra P en unidades estándar es $\dfrac{P-p}{\sigma_P} = \dfrac{P-p}{\sqrt{p(1-p)/n}}$.

Los valores más grandes y más pequeños de esta variable estandarizada son $\pm z_c$, donde z_c determina el nivel de confianza. En estos valores extremos debemos, por lo tanto, tener

$$P - p = \pm z_c\sqrt{\frac{p(1-p)}{n}}$$

Elevando ambos lados al cuadrado,

$$P^2 - 2pP + p^2 = z_c^2\,\frac{p(1-p)}{n}$$

Multiplicando ambos lados por n y simplificando, encontramos

$$(n + z_c^2)p^2 - (2nP + z_c^2)p + nP^2 = 0$$

Si $a = n + z_c^2$, $b = -(2nP + z_c^2)$ y $c = nP^2$, esta ecuación se convierte en $ap^2 + bp + c = 0$, cuya solución para p está dada por la fórmula cuadrática

$$p = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{2nP + z_c^2 \pm \sqrt{(2nP + z_c^2)^2 - 4(n + z_c^2)(nP^2)}}{2(n + z_c^2)}$$

$$= \frac{2nP + z_c^2 \pm z_c\sqrt{4nP(1-P) + z_c^2}}{2(n + z_c^2)}$$

Dividiendo el numerador y denominador por $2n$, se obtiene

$$p = \frac{P + \dfrac{z_c^2}{2n} \pm z_c \sqrt{\dfrac{P(1-P)}{n} + \dfrac{z_c^2}{4n^2}}}{1 + \dfrac{z_c^2}{n}}$$

(*b*) Para límites de confianza del 99.73%, $z_c = 3$. Entonces, usando $P = 0.55$ y $n = 100$ en la fórmula derivada en (*a*), encontramos que $p = 0.40$ y 0.69, lo cual coincide con el problema 6.13(*c*).

(*c*) Si n es grande, entonces $z_c^2/2n$, $z_c^2/4n^2$, y z_c^2/n son despreciablemente pequeños y pueden remplazarse por cero, de modo que se obtiene el resultado requerido.

6.28. ¿Es posible obtener un intervalo de confianza del 95% para la desviación estándar de la población, cuya amplitud es más pequeña que la encontrada en el problema 6.22(*a*)?

Los límites de confianza del 95% para la desviación estándar de la población, como se encontró en el problema 6.22(*a*) se obtuvieron escogiendo valores críticos de χ^2, tales que el área en cada cola fuera 2.5%. Es posible encontrar otros límites de confianza del 95% seleccionando valores críticos de χ^2, para los cuales la suma de las áreas de las colas es 5% o 0.05, pero tales áreas en cada cola no son iguales.

En la tabla 6-2 se obtuvieron varios de esos valores críticos y se muestran los intervalos de confianza del 95% correspondientes.

Tabla 6-2

Valores críticos	Intervalo de confianza del 95%	Amplitud
$\chi_{0.01} = 12.44$, $\chi_{0.96} = 15.32$	92.3 a 113.7	21.4
$\chi_{0.02} = 12.64$, $\chi_{0.97} = 15.42$	91.7 a 111.9	20.2
$\chi_{0.03} = 12.76$, $\chi_{0.98} = 15.54$	91.0 a 110.8	19.8
$\chi_{0.04} = 12.85$, $\chi_{0.99} = 15.73$	89.9 a 110.0	20.1

A partir de esta tabla se ve que un intervalo del 95% con amplitud de sólo 19.8, es de 91.0 a 110.8.

Se puede encontrar un intervalo con una amplitud aún menor si continuamos con el mismo método de aproximación, usando valores críticos tales como $\chi_{0.031}$ y $\chi_{0.981}$, $\chi_{0.032}$ y $\chi_{0.982}$, etc.

Sin embargo, en general la disminución del intervalo que de aquí se obtiene es despreciable, de manera que no se justifica todo el esfuerzo que implica.

Problemas suplementarios

ESTIMACIÓN INSESGADA Y ESTIMACIÓN EFICIENTE

6.29. Se determinaron las medidas de una muestra de pesos como 8.3, 10.6, 9.7, 8.8, 10.2 y 9.4 lb, respectivamente. Determine los estimadores insesgados y eficientes de (*a*) la media de la población, (*b*) la varianza de la población. (*c*) compare la desviación estándar muestral con la desviación estándar estimada de la población.

6.30. Una muestra de 10 tubos de televisión producidos por una compañía mostró un tiempo de vida medio de 1200 horas y desviación estándar de 100 horas. Estime (*a*) la media, (*b*) la desviación estándar de la población de todos los tubos de televisión que produce esa compañía.

6.31. (*a*) Desarrolle el problema 6.30 si se obtienen los mismos resultados para 30, 50 y 100 tubos de televisión. (*b*) ¿Qué concluye usted sobre la relación entre las desviaciones estándar muestrales y las desviaciones estándar de la población para muestras de tamaños diferentes?

ESTIMACIÓN DEL INTERVALO DE CONFIANZA PARA MEDIAS (MUESTRAS GRANDES)

6.32. La media y la desviación estándar de las cargas máximas soportadas por 60 cables (*véase* problema 5.98) son 11.09 toneladas y 0.73 toneladas respectivamente. Encuentre los límites de confianza del (*a*) 95%, (*b*) 99% para la media de las cargas máximas de todos los cables que produce esa compañía.

6.33. La media y la desviación estándar de los diámetros de una muestra de 250 remaches hechos por una compañía son 0.72642 centímetros y 0.00058 centímetros respectivamente (*véase* problema 5.100). Encuentre los límites de confianza (*a*) del 99%, (*b*) 98%, (*c*) 95%, (*d*) 90% para el diámetro promedio de todos los remaches fabricados por esa compañía.

6.34. Encuentre (*a*) los límites de confianza del 50%, (*b*) el error probable para el diámetro promedio del problema 6.33.

6.35. Si la desviación estándar de los tiempos de vida de los tubos de televisión se estiman en 100 horas, ¿qué tan grande debe ser una muestra que se tome, para que se tenga confianza del (*a*) 95%, (*b*) 90%, (*c*) 99%, (*d*) 99.73% de que el error en el tiempo estimado de la vida media no exceda las 20 horas?

6.36. ¿Cuáles son los tamaños de las muestras del problema 6.35 si el error en el estimado del tiempo de vida media no debe exceder las 10 horas?

ESTIMACIÓN DEL INTERVALO DE CONFIANZA PARA MEDIAS (MUESTRAS PEQUEÑAS)

6.37. Una muestra de 12 medidas de resistencia al rompimiento de hebras de algodón dieron una media de 7.30 onzas y desviación estándar de 1.24 onzas. Encuentre los límites de confianza del (*a*) 95%, (*b*) 99% para la resistencia media verdadera al rompimiento.

6.38. Desarrolle el problema 6.37 suponiendo que aplica el método de la teoría para muestras grandes, y compare los resultados obtenidos.

6.39. Se registraron cinco medidas del tiempo de reacción de un individuo ante cierto estímulo como 0.28, 0.30, 0.27, 0.33, 0.31 segundos. Encuentre los límites de confianza del (*a*) 95%, (*b*) 99% para el tiempo verdadero medio de reacción.

ESTIMACIÓN DEL INTERVALO DE CONFIANZA PARA PROPORCIONES

6.40. Una urna contiene canicas rojas y blancas de proporción desconocida. Una muestra aleatoria de 60 canicas seleccionadas con remplazo de la urna mostró que el 70% eran rojas. Encuentre los límites de confianza del (*a*) 95%, (*b*) 99%, (*c*) 99.73% para la proporción verdadera de canicas rojas en la urna. Presente los resultados usando la fórmula aproximada y la fórmula exacta del problema 6.27.

6.41. ¿Qué tan larga debe ser la muestra de canicas que se debe tomar en el problema 6.40 para tener confianza del (*a*) 95%, (*b*) 99%, (*c*) 99.73% de que las proporciones muestrales y verdaderas no difieran en más del 5%?

6.42. Se cree que en las elecciones, los resultados de dos candidatos estarán muy cerca. Explique en términos de un ejemplo, y planteando todos los supuestos, ¿cómo determinaría usted el número mínimo de votantes que deben votar, para tener confianza del (*a*) 80%, (*b*) 95%, (*c*) 99% de una decisión a favor de cualquiera de los candidatos?

INTERVALO DE CONFIANZA PARA DIFERENCIAS Y SUMAS

6.43. Se tienen dos grupos similares de pacientes, *A* y *B*, con 50 y 100 individuos, respectivamente. El grupo *A* recibió un nuevo tipo de píldora para dormir y el segundo, un tipo convencional. Para los pacientes del grupo *A* el número promedio de horas de sueño fue 7.82 con desviación estándar 0.24 horas. Para los pacientes del grupo *B* el número promedio de horas de sueño fue 6.75 con desviación estándar 0.30 horas. Encuentre los límites de

confianza del (*a*) 95% y (*b*) 99% para la diferencia en el número promedio de horas de sueño inducidas por los dos tipos de píldoras para dormir.

6.44. Una muestra de 200 tornillos de una máquina mostró que 15 eran defectuosos, mientras otra de 100 tornillos, de otra máquina, mostró que 12 eran defectuosos. Encuentre los límites de confianza del (*a*) 95%, (*b*) 99%, (*c*) 99.73% para la diferencia en la proporción de tornillos defectuosos de las dos máquinas. Discuta los resultados obtenidos.

6.45. Una compañía fabrica balineras con peso promedio de 0.638 onzas y desviación estándar 0.012 onzas. Encuentre los límites de confianza del (*a*) 95%, (*b*) 99% para los pesos de lotes de 100 balineras cada uno.

INTERVALO DE CONFIANZA PARA VARIANZAS O DESVIACIONES ESTÁNDAR

6.46. La desviación estándar de la resistencia al rompimiento de 100 cables probados por una compañía fue 1800 lb. Encuentre los límites de confianza del (*a*) 95%, (*b*) 99%, (*c*) 99.73% para la desviación estándar de todos los cables producidos por esa compañía.

6.47. ¿Qué tan grande debe ser una muestra para tener confianza del (*a*) 95%, (*b*) 99%, (*c*) 99.73% de que la desviación estándar de la población no diferirá de la desviación estándar muestral por más del 2%?

6.48. La desviación estándar de los tiempos de vida de 10 bombillas eléctricas fabricadas por una compañía es 120 horas. Encuentre los límites de confianza del (*a*) 95%, (*b*) 99% para la desviación estándar de todas las bombillas fabricadas por la compañía.

6.49. Desarrolle el problema 6.48 si 25 bombillas eléctricas muestran la misma desviación estándar de 120 horas.

6.50. Desarrolle el problema 6.48 usando la distribución χ^2 si una muestra de 100 bombillas eléctricas muestran la misma desviación estándar de 120 horas.

INTERVALO DE CONFIANZA PARA EL COCIENTE ENTRE VARIANZAS

6.51. Se encontró que las desviaciones estándar de los diámetros de balineras producidas por dos máquinas eran 0.042 cm y 0.035 cm, respectivamente, con base en muestras de tamaño 10 cada una. Encuentre los intervalos de confianza del (*a*) 98%, (*b*) 90% para el cociente de varianzas.

6.52. Determine los intervalos de confianza del (*a*) 98%, (*b*) 90% para el cociente de las desviaciones estándar del problema 6.51.

6.53. Dos muestras de tamaño 6 y 8, respectivamente, mostraron tener la misma varianza. Encuentre los intervalos de confianza del (*a*) 98%, (*b*) 90% para el cociente de varianzas de las poblaciones de las cuales se obtuvieron.

6.54. *Desarrolle (a) el problema 6.51, (b) el problema 6.53 si cada muestra tuviera tamaño 120.*

ESTIMACIÓN DE MÁXIMA VEROSIMILITUD

6.55. Suponga que se hacen n observaciones, $X_1, \ldots, X_n$, de una distribución de Poisson con parámetro λ desconocido. Encuentre el estimado de máxima verosimilitud de λ.

6.56. Una población tiene función de densidad dada por $f(x) = 2\nu\sqrt{\nu/\pi}\,x^2e^{-\nu x^2}$, $-\infty < x < \infty$. Para n observaciones, $X_1, \ldots, X_n$, hechas de esta población, encuentre el estimador de máxima verosimilitud de ν.

6.57. Una población tiene función de densidad dada por

$$f(x) = \begin{cases} (k+1)x^k & 0 \leqq x \leqq 1 \\ 0 & \text{de otra manera} \end{cases}$$

Para n observaciones, $X_1, \ldots, X_n$, hechas de esta población, encuentre el estimador de máxima verosimilitud de k.

PROBLEMAS DIVERSOS

6.58. Los coeficientes de confianza del 99% (dos colas) para la distribución normal están dados por ± 2.58. ¿Cuáles son los coeficientes correspondientes para la distribución t si (*a*) $\nu = 4$, (*b*) $\nu = 12$, (*c*) $\nu = 25$, (*d*) $\nu = 30$, (*e*) $\nu = 40$?

6.59. Una compañía tiene 500 cables. Una prueba de 40 cables seleccionados al azar mostró una media de resistencia al rompimiento de 2400 lb y desviación estándar 150 lb. (*a*) ¿Cuáles son los límites de confianza del 95% y 99% para estimar la resistencia media al rompimiento de los 460 cables restantes? (*b*) ¿Con cuál grado de confianza podemos decir que la resistencia promedio al rompimiento de los 460 cables restantes es 2400 ± 35 lb?

Respuestas a los problemas suplementarios

6.29. (*a*) 9.5 lb (*b*) 0.74 lb^2 (*c*) 0.78 y 0.86 lb, respectivamente

6.30. (*a*) 1200 horas (*b*) 105.4 horas

6.31. (*a*) Los estimadores de las desviaciones estándar de la población para tamaños muestrales de 30, 50 y 100 tubos son, respectivamente, 101.7, 101.0 y 100.5 horas. En todos los casos los estimadores de los medios de la población son 1200 horas.

6.32. (*a*) 11.09 ± 0.18 toneladas (*b*) 11.09 ± 0.24 toneladas

6.33. (*a*) 0.72642 ± 0.000095 centímetros (*c*) 0.72642 ± 0.000072 centímetros
(*b*) 0.72642 ± 0.000085 centímetros (*d*) 0.72642 ± 0.000060 centímetros

6.34. (*a*) 0.72642 ± 0.000025 centímetros (*b*) 0.000025 centímetros

6.35. (*a*) al menos 97 (*b*) al menos 68 (*c*) al menos 167 (*d*) al menos 225

6.36. (*a*) al menos 385 (*b*) al menos 271 (*c*) al menos 666 (*d*) al menos 900

6.37. (*a*) 7.38 ± 0.82 onzas (*b*) 7.38 ± 1.16 onzas

6.38. (*a*) 7.38 ± 0.70 onzas (*b*) 7.38 ± 0.96 onzas

6.39. (*a*) 0.298 ± 0.030 segundos (*b*) 0.298 ± 0.049 segundos

6.40. (*a*) 0.70 ± 0.12, 0.69 ± 0.11 (*b*) 0.70 ± 0.15, 0.68 ± 0.15 (*c*) 0.70 ± 0.18, 0.67 ± 0.17

6.41. (*a*) al menos 323 (*b*) al menos 560 (*c*) al menos 756

6.43. (*a*) 1.07 ± 0.09 horas (*b*) 1.07 ± 0.12 horas

6.44. (*a*) 0.045 ± 0.073 (*b*) 0.045 ± 0.097 (*c*) 0.045 ± 0.112

6.45. (*a*) 63.8 ± 0.24 onzas (*b*) 63.8 ± 0.31 onzas

6.46. (*a*) 1800 ± 249 lb (*b*) 1800 ± 328 lb (*c*) 1800 ± 382 lb

6.47. (*a*) al menos 4802 (*b*) al menos 8321 (*c*) al menos 11,250

6.48. (*a*) 87.0 a 230.9 horas (*b*) 78.1 a 288.5 horas

6.49. (*a*) 95.6 a 170.4 horas (*b*) 88.9 a 190.8 horas

6.50. (*a*) 106.1 a 140.5 horas (*b*) 102.1 a 148.1 horas

6.51. (*a*) 0.269 a 7.70 (*b*) 0.453 a 4.58

6.52. (*a*) 0.519 a 2.78 (*b*) 0.673 a 2.14

6.53. (*a*) 0.140 a 11.025 (*b*) 0.264 a 5.124

6.54. (*a*) 0.941 a 2.20, 1.067 a 1.944 (*b*) 0.654 a 1.53, 0.741 a 1.35

6.55. $\lambda = (\sum x_k)/n$

6.56. $\nu = \dfrac{3n}{2(x_1^2 + \cdots + x_n^2)}$

6.57. $k = -1 - \dfrac{n}{\ln (x_1 \cdots x_n)}$

6.58. (*a*) ±4.60 (*b*) ±3.06 (*c*) ±2.79 (*d*) ±2.75 (*e*) ±2.70

6.59. (*a*) 2400 ± 45 lb, 2400 ± 59 lb (*b*) 87.6%

CAPÍTULO 7

Pruebas de hipótesis y significancia

DECISIONES ESTADÍSTICAS

Muy a menudo, en la práctica, debemos tomar decisiones sobre poblaciones, con base en la información que nos da la muestra. Tales decisiones se conocen como *decisiones estadísticas*. Por ejemplo, podemos querer decidir a partir de datos muestrales si un suero nuevo es realmente efectivo para curar una enfermedad, si un procedimiento educativo es mejor que otro, o si cierta moneda está cargada.

HIPÓTESIS ESTADÍSTICAS. HIPÓTESIS NULA

Con el fin de llegar a una decisión es útil hacer suposiciones o conjeturas sobre las poblaciones involucradas. Tales suposiciones, que pueden ser o no ciertas, reciben el nombre de *hipótesis estadísticas* y, en general, son planteamientos sobre la distribución de probabilidad de las poblaciones.

Por ejemplo, si vamos a decidir si cierta moneda está cargada, formulamos la hipótesis de que la moneda es balanceada, es decir, $p = 0.5$, donde p es la probabilidad de obtener caras. De igual manera, si queremos decidir si un procedimiento es mejor que otro, formulamos la hipótesis de que *no hay diferencia* entre los procedimientos (así, cualquier diferencia observada se debe a meras fluctuaciones en el muestreo de la misma población). Tales hipótesis se llaman *hipótesis nulas* o simplemente *hipótesis*, y se denotan como H_0.

Cualquier hipótesis que difiera de la hipótesis nula dada se llama *hipótesis alterna*. Por ejemplo, si la hipótesis nula es $p = 0.5$, las hipótesis alternas posibles son $p = 0.7$, $p \neq 0.5$ o $p > 5$. Una hipótesis alterna de la hipótesis nula se denota por H_1.

PRUEBAS DE HIPÓTESIS Y SIGNIFICANCIA

Si suponemos que cierta hipótesis es verdadera, pero encontramos que los resultados de una muestra aleatoria difieren marcadamente de los esperados bajo la hipótesis sobre la base del azar de la teoría del muestreo, podremos decir que las diferencias observadas son *significativas* y nos inclinamos a rechazar la

hipótesis (o al menos a no aceptarla debido a la evidencia obtenida). Por ejemplo, si de 20 lanzamientos de una moneda obtenemos 16 caras, podemos inclinarlos a rechazar la hipótesis de que la moneda es balanceada, aunque posiblemente estemos equivocados.

Los procedimientos que nos permiten aceptar o rechazar hipótesis o determinar si las muestras observadas difieren significativamente de los resultados esperados reciben el nombre de *pruebas de hipótesis*, *pruebas de significancia* o *reglas de decisión*.

ERRORES DE TIPO I Y TIPO II

Si rechazamos una hipótesis cuando da la casualidad que es verdadera, decimos que se ha cometido un *error de tipo I*. Si, por el contrario, aceptamos una hipótesis cuando ésta ha debido rechazarse, decimos que se ha cometido un *error de tipo II*. En cualquier caso, se ha tomado una mala decisión o se ha hecho un error de juicio.

Para que cualquier prueba de hipótesis o las reglas de decisión sean adecuadas, se deben diseñar de manera que reduzcan los errores de la decisión. Esto no es sencillo ya que, para un tamaño muestral dado, el intento de disminuir un tipo de error va generalmente acompañado del incremento en el otro tipo de error. En la práctica, un tipo de error puede ser más grave que el otro, de manera que debemos inclinarnos a favor de una limitación del error más grave. La única manera de reducir ambos tipos de error es incrementando el tamaño de la muestra, lo cual puede o no ser posible.

NIVEL DE SIGNIFICANCIA

Al probar una hipótesis dada, la probabilidad máxima con la que queremos tomar el riesgo de un error tipo I se llama *nivel de significancia* de la prueba. Esta probabilidad se especifica antes de que se hayan tomado muestras, para que los resultados obtenidos no influyan en nuestra decisión.

En la práctica, niveles de significancia de 0.05 ó 0.01 son habituales, aunque se pueden usar otros valores. Si por ejemplo, al diseñar una prueba de hipótesis escogemos un nivel de significancia de 0.05 ó 5%, entonces hay una posibilidad de 5 en 100 de que rechacemos la hipótesis cuando deberíamos aceptarla, es decir, siempre que la hipótesis nula sea verdadera, tenemos cerca del 95% de confianza de que tomaremos la decisión correcta. En tales casos decimos que rechazamos la hipótesis con un nivel de significancia de 0.05, lo cual significa que podemos estar equivocados con probabilidad de 0.05.

PRUEBAS QUE INVOLUCRAN LA DISTRIBUCIÓN NORMAL

Para ilustrar las ideas presentadas anteriormente, supongamos que bajo cierta hipótesis la distribución muestral del estadístico S tiene distribución normal con media μ_S y desviación estándar σ_S. De igual manera, supongamos que decidimos rechazar la hipótesis si S es muy grande o muy pequeña. La distribución de la variable estandarizada $Z = (S - \mu_S)/\sigma_S$ es la distribución normal (media 0, varianza 1), que se muestra en la figura 7-1, y los valores extremos de Z pueden llevar al rechazo de la hipótesis.

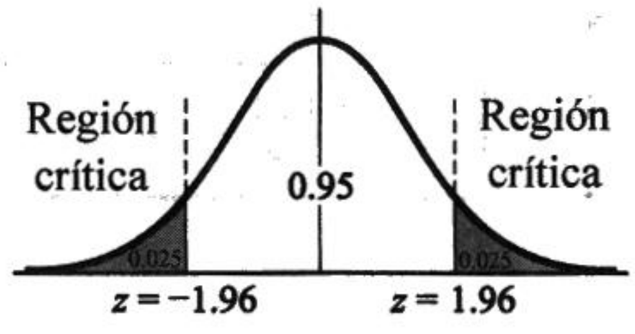


Fig. 7-1

Como se indica en la figura, podemos tener confianza del 95% de que, si la hipótesis es verdadera, el valor z de un estadístico muestral verdadero S estará entre -1.96 y 1.96 (ya que el área bajo la curva normal entre estos valores es 0.95).

Sin embargo, si al obtener una muestra única al azar encontramos que el valor z de ese estadístico está fuera del rango -1.96 a 1.96, podremos concluir que tal evento es susceptible de ocurrir con probabilidad de sólo 0.05 (área total sombreada de la figura) si la hipótesis dada es verdadera. Podemos, entonces, decir que el valor z difiere significativamente de lo esperado bajo la hipótesis, e inclinarnos a rechazarla.

El área total sombreada 0.05 es el nivel de significancia de la prueba, y representa la probabilidad de que estemos equivocados al rechazar la hipótesis, es decir, la probabilidad de que ocurra un error de tipo I. Por consiguiente, decimos que la hipótesis fue *rechazada con un nivel de significancia* de 0.05 o que el valor z del estadístico muestral dado *es significante con un nivel de significancia* del 0.05.

El conjunto de valores z por fuera del rango -1.96 a 1.96 constituye lo que llamamos la *región crítica* o *región de rechazo de la hipótesis* o la *región de significancia*. El conjunto de valores z ubicados dentro del rango -1.96 a 1.96 puede, entonces, llamarse *región de aceptación de la hipótesis* o *región de no significancia*.

A partir de las observaciones anteriores podemos formular la siguiente regla de decisión:

(*a*) Rechazar la hipótesis con un nivel de significancia del 0.05 si el valor z del estadístico S está fuera del rango -1.96 a 1.96 (es decir, si $z > 1.96$ o $z < 1.96$). Esto equivale a decir que el estadístico muestral observado es significante al nivel del 0.05.

(*b*) Aceptar la hipótesis (o, si se desea, no tomar una decisión).

Se debe notar que se pueden emplear otros niveles de significancia. Por ejemplo, si se usara un nivel de 0.01, podríamos remplazar 1.96 visto anteriormente por 2.58 (*véase* tabla 7-1). También se puede usar la tabla 6-1, página 210, ya que la suma del nivel de significancia y del nivel de confianza es 100%.

PRUEBAS DE UNA Y DOS COLAS

En la prueba anterior nos interesamos en los valores extremos del estadístico S o en su correspondiente valor z a ambos lados de la media, es decir, en ambas colas de la distribución. Por esta razón, tales pruebas reciben el nombre de *pruebas de dos colas* o *pruebas bilaterales*.

Sin embargo, con frecuencia podemos estar interesados sólo en valores extremos de un lado de la media, es decir, en una cola de la distribución, como, por ejemplo, cuando estamos probando la hipótesis de que un proceso es mejor que otro (lo cual es diferente de probar si un proceso es mejor o peor que otro). Tales pruebas reciben el nombre de *pruebas de una cola* o *pruebas unilaterales*. En tales casos la región crítica es una región a un lado de la distribución, con un área igual al nivel de significancia.

La tabla 7-1, que da valores críticos de z para pruebas de dos colas y una cola a varios niveles de significancia, será útil para propósitos de referencia. Valores críticos de z para otros niveles de significancia se encuentran en la tabla de las áreas de la curva normal.

Tabla 7-1

Nivel de significancia α	0.10	0.05	0.01	0.005	0.002
Valores críticos de z para pruebas de una cola	-1.28 *o* 1.28	-1.645 *o* 1.645	-2.33 *o* 2.33	-2.58 *o* 2.58	-2.88 *o* 2.88
Valores críticos de z para pruebas de dos colas	-1.645 *y* 1.645	-1.96 *y* 1.96	-2.58 *y* 2.58	-2.81 *y* 2.81	-3.08 *y* 3.08

EL VALOR *P*

En la mayoría de las pruebas que tengamos en cuenta, la hipótesis nula H_0 será una afirmación de que cierto parámetro de una población tiene un valor específico, y la hipótesis alterna H_1 será alguna de las siguientes afirmaciones:

(*i*) El parámetro es mayor que el valor indicado (prueba de cola derecha).

(*ii*) El parámetro es menor que el valor indicado (prueba de cola izquierda).

(*iii*) El parámetro es mayor o menor que el valor indicado (prueba de dos colas).

En los casos (*i*) y (*ii*), H_1 tiene dirección única con respecto al parámetro, y en el caso (*iii*) H_1 es bidireccional. Después de que se lleva a cabo la prueba y se calcula el estadístico S, el valor P de la prueba es la probabilidad de que un valor S en la dirección de H_1 y tan extremo como el valor obtenido se obtenga si H_0 fuera verdadero.

Por ejemplo, supongamos que la desviación estándar σ de una población normal es 3, y H_0 afirma que la media μ es igual a 12. De una población se obtiene una muestra aleatoria de tamaño 36, y se obtiene una media muestral $\bar{x} = 12.95$. La prueba estadística es

$$Z = \frac{\bar{X} - 12}{\sigma/\sqrt{n}} = \frac{\bar{X} - 12}{0.5},$$

la cual, si H_0 es verdadera, es la variable aleatoria normal estándar. El valor de Z en la prueba es $(12.95 - 12)/0.5 = 1.9$. El valor P para la prueba depende entonces de la hipótesis alterna H_1 como sigue:

(*i*) Para H_1: $\mu > 12$ [caso (*i*) anterior], el valor P es la probabilidad de que de una muestra aleatoria de tamaño 36 se obtenga una media muestral de 12.95 o más, si la media verdadera es 12, es decir, $P(Z \geq 1.9) = 0.029$. En otras palabras, las oportunidades de que $\bar{x} > 12.95$ si $\mu = 12$ son 3 en 100.

(*ii*) Para H_1: $\mu < 12$ [caso (*ii*) anterior], el valor P es la probabilidad de que de una muestra aleatoria de tamaño 36 se obtenga una media muestral de 12.95 o menos, si la media verdadera es 12, es decir, $P(Z \leq 1.9) = 0.97$, o las oportunidades de que $\bar{x} \leq 12.95$ si $\mu = 12$ son alrededor de 97 en 100.

(*iii*) Para H_1: $\mu \neq 12$ [caso (*iii*) anterior], el valor P es la probabilidad de que de una muestra aleatoria de tamaño 36 se obtenga una media muestral de 0.95 o más unidades lejos de 12, es decir, $\bar{x} \geq 12.95$ o ≤ 11.05, si la media verdadera es 12. Aquí el valor P es $P(Z \geq 1.9) + P(Z \leq -1.9) = 0.057$, lo cual indica que las probabilidades de que $|\bar{x} - 12| \geq 0.95$ si $\mu = 12$ son alrededor de 6 en 100.

Los valores pequeños de P proveen evidencia para rechazar la hipótesis nula a favor de la hipótesis alterna, y valores grandes de P dan evidencia para no rechazar la hipótesis nula a favor de una hipótesis alterna. En el caso (*i*) anterior, el valor pequeño P, 0.029, es un indicador bastante fuerte de que la media de la población es mayor que 12, mientras que en el caso (*ii*) el valor grande de P, 0.97 sugiere que no se debe rechazar H_0: $\mu = 12$ a favor de H_1: $\mu < 12$. En el caso (*iii*) el valor P, 0.057 da evidencia para rechazar H_0 a favor de H_1: $\mu \neq 12$, pero no tanta como se da para rechazar H_0 a favor de H_1: $\mu > 12$.

Debemos tener presente que el valor P y el nivel de significancia no proveen criterios para rechazar o no la hipótesis nula por sí misma, sino para rechazarla o no a favor de una hipótesis alterna. Tal como lo ilustra el ejemplo anterior, los resultados idénticos de las pruebas y niveles de significancia pueden conducir a conclusiones diferentes, considerando la misma hipótesis nula en relación con diferentes hipótesis alternas.

Cuando el estadístico S es la variable aleatoria normal estándar, la tabla del apéndice C es suficiente para calcular el valor P, pero cuando S es una de las variables aleatorias t, F o chi cuadrado, todas con distribuciones diferentes dependiendo de sus grados de libertad, es necesario utilizar software especializado de computador o tablas más extensas que las de los apéndices D, E y F, para calcular el valor P.

PRUEBAS ESPECIALES DE SIGNIFICANCIA PARA MUESTRAS GRANDES

Para muestras grandes, muchos estadísticos S tienen distribuciones casi normales con media μ_S y desviación estándar σ_S. En tales casos podemos usar los resultados anteriores para formular reglas de decisión para las pruebas de hipótesis y significancia. Los siguientes casos especiales son sólo unos pocos de los estadísticos de interés práctico. En cada caso los resultados son válidos para poblaciones infinitas o para muestreo

con remplazo. Los resultados deben modificarse para la toma de muestras sin remplazo de poblaciones finitas.

1. **MEDIAS.** Aquí $S = \bar{X}$, la media muestral; $\mu_s = \mu_{\bar{X}} = \mu$, la media de la población; $\sigma_S = \sigma_{\bar{X}} = \sigma/\sqrt{n}$, donde σ es la desviación estándar de la población y n es el tamaño muestral. La variable estandarizada está dada por

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{1}$$

Cuando es necesario, la desviación estándar muestral, s (o $\hat{s}$), se utiliza para estimar σ.

Para probar la hipótesis nula H_0 de que la media de la población es $\mu = a$ podemos usar el estadístico (*1*). Entonces, si la hipótesis alterna es $\mu \neq a$, usando la prueba de dos colas, podemos aceptar H_0 (o al menos no rechazarla) al nivel de significancia de 0.05 si para una muestra particular de tamaño n con media $\bar{x}$

$$-1.96 \leqq \frac{\bar{x} - a}{\sigma/\sqrt{n}} \leqq 1.96 \tag{2}$$

y podemos rechazarla de otra manera. Para otros niveles de significancia podemos cambiar (*2*) de manera apropiada. Para probar H_0 en contra de la hipótesis alterna de que la media de la población es mayor que a, podemos usar la prueba de una cola y aceptar H_0 (o al menos no rechazarla) al nivel del 0.05 si

$$\frac{\bar{x} - a}{\sigma/\sqrt{n}} < 1.645 \tag{3}$$

(*véase* tabla 7-1) y rechazarla de otra manera. Para probar H_0 en contra de la hipótesis alterna de que la media de la población es menor que a, podemos aceptar H_0 al nivel del 0.05 si

$$\frac{\bar{x} - a}{\sigma/\sqrt{n}} > -1.645 \tag{4}$$

2. **PROPORCIONES.** Aquí $S = P$, la proporción de "éxitos" en una muestra; $\mu_S = \mu_p = p$, donde p es la proporción de éxitos en la población y n es el tamaño muestral; $\sigma_S = \sigma_P = \sqrt{pq/n}$, donde $q = 1 - p$. La variable estandarizada está dada por

$$Z = \frac{P - p}{\sqrt{pq/n}} \tag{5}$$

En el caso $P = X/n$, donde X es el número verdadero de éxitos en una muestra, (*5*) se convierte en

$$Z = \frac{X - np}{\sqrt{npq}} \tag{6}$$

Se pueden hacer observaciones similares a las hechas atrás sobre pruebas de una y dos colas para medias.

3. **DIFERENCIA DE MEDIAS.** Sean $\bar{X}_1$ y $\bar{X}_2$ las medias muestrales obtenidas en muestras grandes de tamaños n_1 y n_2 de poblaciones respectivas con media μ_1 y μ_2 y desviaciones estándar σ_1 y σ_2. Considere la hipótesis nula de que *no hay diferencia* entre las medias de las poblaciones, es decir, $\mu_1 = \mu_2$. De (*11*), página 169, remplazando $\mu_1 = \mu_2$ vemos que la distribución muestral de diferencias en medias es aproximadamente normal con media y desviación estándar dadas por

$$\mu_{\bar{X}_1-\bar{X}_2} = 0 \qquad \sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{7}$$

donde podemos, si es necesario, usar las desviaciones estándar observadas s_1 y s_2 (o $\hat{s}_1$ y $\hat{s}_2$) como estimados de σ_1 y σ_2.

Usando la variable estandarizada dada por

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - 0}{\sigma_{\bar{X}_1-\bar{X}_2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1-\bar{X}_2}} \tag{8}$$

de manera similar a la descrita en la parte 1, atrás, podemos probar la hipótesis nula en contra de la hipótesis alterna (o de la significancia de la diferencia observada) a un nivel adecuado de significancia.

4. **DIFERENCIA DE PROPORCIONES.** Sean P_1 y P_2 las proporciones muestrales obtenidas en muestras grandes de tamaños n_1 y n_2 de las poblaciones respectivas con proporciones p_1 y p_2. Considere la hipótesis nula de que no hay diferencia entre las proporciones de la población, es decir, $p_1 = p_2$, y por lo tanto, de que las muestras se tomaron realmente de la misma población.

De (*13*), página 170, remplazando $p_1 = p_2 = p$, vemos que la distribución muestral de diferencias en proporciones es aproximadamente normal con media y desviación estándar dadas por

$$\mu_{P_1-P_2} = 0 \qquad \sigma_{P_1-P_2} = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{9}$$

donde $\bar{P} = \dfrac{n_1P_1 + n_2P_2}{n_1 + n_2}$ se usa como un estimado de la proporción de la población p.

Usando la variable estandarizada

$$Z = \frac{P_1 - P_2 - 0}{\sigma_{P_1-P_2}} = \frac{P_1 - P_2}{\sigma_{P_1-P_2}} \tag{10}$$

podemos probar diferencias observadas a un nivel apropiado de significancia y de allí probar la hipótesis nula.

De manera similar, se pueden diseñar pruebas que involucren otros estadísticos (*véase tabla 5-1*, página 172).

PRUEBAS ESPECIALES DE SIGNIFICANCIA PARA MUESTRAS PEQUEÑAS

En el caso de que las muestras sean pequeñas ($n < 30$), podemos formular pruebas de hipótesis y significancia usando otras distribuciones además de la normal, tales como la t de Student, chi cuadrado y F. Éstas involucran teoría de muestras correctas y así, por supuesto, se mantienen válidas incluso cuando las muestras son grandes, en cuyo caso se reducen a aquéllas dadas atrás. Los siguientes son algunos ejemplos.

1. **MEDIAS.** Para probar la hipótesis H_0 de que una población tiene media, μ, usamos

$$T = \frac{\bar{X} - \mu}{S}\sqrt{n-1} = \frac{\bar{X} - \mu}{\hat{S}}\sqrt{n} \tag{11}$$

donde $\bar{X}$ es la media de la muestra de tamaño n. Esto es análogo usar la variable estandarizada $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ para n grande, excepto que $\hat{S} = \sqrt{n/(n-1)}\ S$ se usa en lugar de σ. La diferencia radica en que mientras

Z tiene distribución normal, T tiene distribución t de Student. A medida que n se incrementa, esto tiende a concordar. Se pueden hacer pruebas de hipótesis similares a las de las medias de la página 232, usando valores críticos de t en lugar de valores críticos de z.

2. DIFERENCIA DE MEDIAS. Supongamos que se obtienen dos muestras aleatorias de tamaños n_1 y n_2 de poblaciones normales (o aproximadamente normales), cuyas desviaciones estándar son iguales, es decir, $\sigma_1 = \sigma_2$. Supongamos, además, que estas dos muestras tienen medias y desviaciones estándar dadas por $\bar{X}_1$, $\bar{X}_2$ y s_1, s_2, respectivamente. Para probar la hipótesis H_0 de que las muestras vienen de la misma población (es decir, $\mu_1 = \mu_2$ así como $\sigma_1 = \sigma_2$), usamos la variable dada por

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{donde} \quad \sigma = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}} \tag{12}$$

La distribución de T es la distribución t de Student con $\nu = n_1 + n_2$ grados de libertad. El uso de (*12*) se hace factible al remplazar $\sigma_1 = \sigma_2 = \sigma$ en (*12*), página 169, y luego usarlo como un estimador de σ_2 del promedio ponderado

$$\frac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}$$

donde $\hat{S}_1^2$ y $\hat{S}_2^2$ son los estimadores insesgados de σ_1^2 y σ_2^2. Ésta es la *varianza combinada* obtenida al combinar los datos.

3. VARIANZAS. Para probar la hipótesis H_0 de que una población normal tiene varianza σ^2, consideramos las variables aleatorias

$$\chi^2 = \frac{nS^2}{\sigma^2} = \frac{(n-1)\hat{S}^2}{\sigma^2} \tag{13}$$

la cual (*véase* página 171) tiene la distribución chi cuadrado con $n - 1$ grados de libertad. Entonces, si una muestra aleatoria de tamaño n resulta tener varianza s^2, podemos, sobre la base de la prueba de dos colas, aceptar H_0 (o al menos no rechazarla) al nivel de 0.05 si

$$\chi^2_{0.025} \leqq \frac{ns^2}{\sigma^2} \leqq \chi^2_{0.975} \tag{14}$$

y rechazarla de otra manera. Se obtiene un resultado similar para el nivel de 0.01 u otro nivel.

Para probar la hipótesis H_1 de que la varianza de la población es mayor que σ^2, podemos aún usar la hipótesis nula H_0, pero ahora emplearemos la prueba de una cola. Así, podemos rechazar H_0 al nivel de 0.05 (y por consiguiente concluir que H_1 es correcta) si la varianza s^2 de la muestra particular es tal que

$$\frac{ns^2}{\sigma^2} > \chi^2_{0.95} \tag{15}$$

y aceptamos H_0 (o al menos no la rechazamos) de otra manera.

4. COCIENTE DE VARIANZAS. En algunos problemas queremos decidir si dos muestras de tamaños m y n, respectivamente, cuyas varianzas son s_1^2 y s_2^2, provienen o no de poblaciones normales con la misma varianza. En tales casos, usamos el estadístico (*véase* página 172).

$$F = \frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2} \tag{16}$$

donde σ_1^2 y σ_2^2 son las varianzas de dos poblaciones normales de las cuales se tomaron las muestras. Supongamos que H_0 denota la hipótesis nula de que no hay diferencia entre las varianzas de las poblaciones, es decir $\sigma_1^2 = \sigma_2^2$. Entonces, bajo esta hipótesis (*16*) se convierte en

$$F = \frac{\hat{S}_1^2}{\hat{S}_2^2} \tag{17}$$

Para probar esta hipótesis al nivel de 0.10, por ejemplo, primero debemos observar que F en (*16*) tiene la distribución F con $m - 1, n - 1$ grados de libertad. Entonces, usando una prueba de dos colas, podemos aceptar H_0 (o no rechazarla) al nivel del 0.10 si

$$F_{0.05} \leqq \frac{\hat{s}_1^2}{\hat{s}_2^2} \leqq F_{0.95} \tag{18}$$

y rechazarla de otra manera.

Se pueden formular procedimientos similares usando pruebas de una cola en caso de que queramos probar la hipótesis de que la varianza de una población particular es, de hecho, mayor que la otra.

RELACIÓN ENTRE LA TEORÍA DE LA ESTIMACIÓN Y LA PRUEBA DE HIPÓTESIS

A partir de las anotaciones anteriores se puede observar que existe una relación entre la teoría de la estimación que involucra los intervalos de confianza y la teoría de la prueba de hipótesis. Por ejemplo, observamos que el resultado (*2*) para aceptar la H_0 al nivel del 0.05 es equivalente al resultado (*1*) de la página 210, que conduce al intervalo de confianza del 95%

$$\bar{x} - \frac{1.96\sigma}{\sqrt{n}} \leqq \mu \leqq \bar{x} + \frac{1.96\sigma}{\sqrt{n}} \tag{19}$$

Así, al menos en el caso de las pruebas de dos colas, podemos en realidad emplear los intervalos de confianza del capítulo 6 para la prueba de hipótesis. Un resultado similar para pruebas de una cola requiere intervalos de confianza de una cola (*véase* problema 6.14).

CURVAS DE OPERACIÓN CARACTERÍSTICAS. POTENCIA DE UNA PRUEBA

Hemos visto que el error de tipo I puede limitarse si se escoge adecuadamente un nivel de significancia. Es posible evitar el riesgo de cometer los errores de tipo II, simplemente no cometiéndolos, lo cual significa no aceptar nunca la hipótesis. Sin embargo, en muchos casos prácticos esto no se puede hacer. En tales casos se hace uso de *curvas de operación características*, o *curvas OC*, las cuales son gráficas que muestran las probabilidades de errores de tipo II bajo diferentes hipótesis. Éstas proveen indicaciones de qué tanto ciertas pruebas nos permiten minimizar los errores tipo II, es decir, indican el *poder de una prueba* para evitar que se cometan errores en la toma de decisiones. Son muy útiles para diseñar experimentos mostrando, por ejemplo, los tamaños muestrales que se deben usar.

GRÁFICOS DE CONTROL DE CALIDAD

En la práctica es importante conocer si un proceso ha cambiado lo suficiente como para tomar medidas correctivas. Por ejemplo, tales problemas surgen en el *control de calidad* donde se debe, con rapidez, decidir

si los cambios que se observan obedecen a simples fluctuaciones o a cambios reales en un proceso de fabricación, causados por el deterioro de algunas partes de la máquina o a errores cometidos por los empleados. Los *gráficos de control* proveen un método útil para tratar con dichos problemas (*véase* problema 7.29).

AJUSTE DE LAS DISTRIBUCIONES TEÓRICAS A DISTRIBUCIONES DE FRECUENCIA MUESTRALES

Cuando se tiene cierta indicación de la distribución de una población por razonamiento probabilístico u otro, es posible ajustar tales distribuciones teóricas (también llamadas distribuciones "modelo" o "esperadas") a distribuciones de frecuencia obtenidas de la muestra de una población. El método usado generalmente consiste en emplear la media y la desviación estándar de la muestra para estimar la media y la desviación estándar de la población. *Véanse* problemas 7.30, 7.32 y 7.33.

El problema de probar la *bondad de ajuste* de distribuciones teóricas a distribuciones muestrales es esencialmente el mismo que decidir si hay diferencias significativas entre poblaciones y valores muestrales. A continuación se describe una prueba importante de significancia para la bondad de ajuste de distribuciones teóricas, la *prueba chi cuadrado*.

Con el fin de determinar si una distribución normal representa un buen ajuste para los datos dados, conviene usar el *papel gráfico de curva normal* o *papel gráfico de probabilidad*, como a veces se lo llama (*véase* problema 7.31).

LA PRUEBA DE CHI CUADRADO PARA LA BONDAD DE AJUSTE

Para determinar si la proporción P de "éxitos" en una muestra de tamaño n, obtenida de una población binomial, difiere significativamente de la proporción de la población p de éxitos, usamos el estadístico dado por (*5*) o (*6*) en la página 232. En este caso sencillo sólo pueden ocurrir dos eventos posibles A_1, A_2, a los cuales llamamos "éxito" y "fracaso" y tienen probabilidades p y $q = 1 - p$, respectivamente.

Un valor muestral particular de la variable aleatoria $\times = nP$ se denomina *frecuencia observada* para el evento A_1, mientras que a np se la llama *frecuencia esperada*.

EJEMPLO 7.1. Si obtenemos una muestra de 100 lanzamientos de una moneda balanceada, de manera que $n = 100$, $p = \frac{1}{2}$, entonces la frecuencia esperada de caras (éxitos) es $\boldsymbol{np = (100)(\frac{1}{2}) = 50}$. La frecuencia observada en la muestra puede, por supuesto, ser diferente.

Una generalización evidente es el caso donde ocurren k posibles eventos $A_1, A_2, \ldots, A_k$, cuyas probabilidades respectivas son $p_1, p_2, \ldots, p_k$. En tales casos tenemos una *población multinomial* (*véase* página 122). Si de esta población obtenemos una muestra de tamaño n, las frecuencias observadas para los eventos $A_1, \ldots, A_k$ se pueden describir como variables aleatorias $X_1, \ldots, X_k$ (cuyos valores específicos $x_1, x_2, \ldots, x_k$ serían las frecuencias observadas para la muestra), mientras que las frecuencias esperadas estarían dadas por $np_1, \ldots, np_k$, respectivamente. Los resultados se indican en la tabla 7-2.

Tabla 7-2

Evento	A_1	A_2	$\cdots$	A_k
Frecuencia observada	x_1	x_2	$\cdots$	x_k
Frecuencia esperada	np_1	np_2	$\cdots$	np_k

EJEMPLO 7.2. Si obtenemos una muestra de 120 lanzamientos de un dado balanceado, de manera que $n = 120$, entonces la probabilidad de las caras 1, 2, . . ., 6 se denotan como $p_1, p_2, \ldots, p_6$, respectivamente, y todas son iguales a 1/6. Por supuesto, las frecuencias esperadas de las diferentes caras que surjan en la muestra pueden ser diferentes.

Una clave para la posible generalización del estadístico (*6*), que puede medir las discrepancias existentes entre las frecuencia observada y la esperada en la tabla 7-2 se obtiene elevando al cuadrado el estadístico (6) y escribiéndolo como

$$Z^2 = \frac{(X - np)^2}{npq} = \frac{(X_1 - np)^2}{np} + \frac{(X_2 - nq)^2}{nq} \tag{20}$$

donde $X_1 = \times$ es la variable aleatoria asociada con "éxito" y $X_2 = n - X_1$ es la variable aleatoria asociada con "fracaso". Observe que nq en (*20*) es la frecuencia esperada de fracaso.

La forma del resultado (*20*) sugiere que una medida de la discrepancia entre la frecuencia esperada y la observada para el caso general se suple con el estadístico

$$\chi^2 = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} + \cdots + \frac{(X_k - np_k)^2}{np_k} = \sum_{j=1}^{k} \frac{(X_j - np_j)^2}{np_j} \tag{21}$$

donde la frecuencia total (es decir, el tamaño de la muestra) es n, de manera que

$$X_1 + X_2 + \cdots + X_k = n \tag{22}$$

Una expresión equivalente a (*21*) es

$$\chi^2 = \sum_{j=1}^{k} \frac{X_j^2}{np_j} - n \tag{23}$$

Si $\chi^2 = 0$, las frecuencias observada y esperada concuerdan exactamente, mientras que si $\chi^2 > 0$, no concuerdan de manera exacta. Entre mayor sea el valor de χ^2, mayor será la discrepancia entre las frecuencias observada y esperada.

Tal como se muestra en el problema 7.62, la distribución muestral de χ^2, como se define en (*21*), es aproximadamente muy cercana a la distribución chi cuadrado [de ahí la escogencia de símbolo en (*21*)] si las frecuencias esperadas np_j son al menos iguales a 5, la aproximación mejora para valores grandes. El número de grados de libertad para esta distribución chi cuadrado está dado por:

(*a*) $\nu = k - 1$ si las frecuencias esperadas pueden calcularse sin tener que estimar los parámetros de la población a partir de un estadístico muestral. Observe que restamos 1 de k debido a las limitaciones de (*22*), las cuales estipulan que si conocemos $k - 1$ de las frecuencias esperadas, se puede determinar la frecuencia restante.

(*b*) $\nu = k - 1 - m$ si las frecuencias esperadas pueden calcularse sólo estimando los parámetros m de la población a partir de estadísticos muestrales.

En la práctica, las frecuencias esperadas se calculan sobre la base de la hipótesis H_0. Si bajo esta hipótesis el valor calculado de χ^2 dado por (*21*) o (*23*) es mayor que algunos valores críticos (tales como $\chi^2_{0.95}$ o $\chi^2_{0.99}$, los cuales son valores críticos con niveles de significancia de 0.05 y 0.01, respectivamente), podemos concluir que la frecuencia observada difiere *significativamente* de las frecuencias esperadas y rechazamos H_0, en un procedimiento denominado la *prueba de hipótesis o significancia chi cuadrado*.

Además de aplicar la distribución multinomial, la prueba chi cuadrado se puede usar para determinar qué tan bien otras distribuciones, como la normal o de Poisson, encajan en distribuciones empíricas, es decir, aquellas que se obtienen de datos muestrales (*véase* problema 7.44).

TABLAS DE CONTINGENCIA

La tabla 7-2, en la que las frecuencias observadas ocupan una sola fila, se denomina *tabla de clasificación simple*. Ya que el número de columnas es k, también recibe el nombre de tabla $1 \times k$ (léase "1 por k").

Ampliando estas ideas podemos obtener tablas de clasificación doble o $h \times k$, en las que la frecuencia observada ocupa h filas y k columnas. Dichas tablas se denominan *tablas de contingencia.*

Correspondiendo a cada frecuencia observada en una tabla de contingencia $h \times k$, hay una *frecuencia teórica* o *esperada*, cuyo cálculo está sujeto a ciertas hipótesis según reglas de probabilidad. Estas frecuencias que ocupan las celdas de una tabla de contingencia se denominan *frecuencias elementales*. La frecuencia total en cada fila o en cada columna se llama *frecuencia marginal.*

Para investigar la concordancia entre las frecuencias observada y esperada, calculamos el estadístico

$$\chi^2 = \sum_j \frac{(X_j - np_j)^2}{np_j} \tag{24}$$

donde la suma se extiende en todas las celdas de la tabla de contingencia y los símbolos X_j y np_j representan, respectivamente, las frecuencias observada y esperada en la j-ésima celda. Esta suma, que es análoga a (*21*), contiene hk términos. La suma de todas las frecuencias observadas se denota como n y es igual a la suma de todas las frecuencias esperadas [compare con la ecuación (*22*)].

Como antes, el estadístico (*24*) tiene una distribución muestral dada muy cercanamente por la distribución chi cuadrado, siempre y cuando las frecuencias esperadas no sean muy pequeñas. El número de grados de libertad ν de esta distribución chi cuadrado está dado por $h > 1$, $k > 1$

(*a*) $\nu = (h - 1)(k - 1)$ si las frecuencias esperadas se pueden calcular sin tener que estimar los parámetros de la población para el estadístico muestral. Para una demostración de esto, *véase* el problema 7.48.

(*b*) $\nu = (h - 1)(k - 1) - m$ si las frecuencias esperadas se pueden calcular sólo con estimar m parámetros de la población del estadístico muestral.

Las pruebas de significancia para las tablas $h \times k$ son similares a aquéllas para las tablas $1 \times k$. Se ha encontrado que las frecuencias esperadas están sujetas a hipótesis H_0 particulares. Una hipótesis que se prueba comúnmente quiere decir que las dos clasificaciones son independientes entre sí.

Las tablas de contingencia se pueden extender a un número mayor de dimensiones. Por ejemplo, podemos tener tablas $h \times k \times l$, donde hay 3 clasificaciones.

CORRECCIÓN DE YATES POR LA CONTINUIDAD

Cuando los datos para las distribuciones continuas se aplican a datos discretos, se pueden hacer ciertas correcciones por la continuidad, como vimos en capítulos anteriores. Cuando se usa la distribución chi cuadrado, está disponible una corrección similar. La corrección consiste en reescribir (*21*) como

$$\chi^2 \text{ (corregida)} = \frac{(|X_1 - np_1| - 0.5)^2}{np_1} + \frac{(|X_2 - np_2| - 0.5)^2}{np_2} + \cdots + \frac{(|X_k - np_k| - 0.5)^2}{np_k} \tag{25}$$

lo que se conoce como la *corrección de Yates*. También hay una modificación análoga para (*24*).

En general, la corrección se hace sólo cuando el número total de grados de libertad es $\nu = 1$. Para muestras grandes se obtienen prácticamente los mismos resultados que para la χ^2 no corregida, pero pueden surgir dificultades cerca de valores críticos (*véase* problema 7.41). Para muestras pequeñas donde cada frecuencia esperada está entre 5 y 10, es tal vez mejor comparar tanto los valores corregidos de χ^2 como los no corregidos. Si ambos valores llevan a la misma conclusión con respecto a la hipótesis, como el rechazo al nivel de 0.05, rara vez se presentan dificultades. Si conducen a diferentes conclusiones, se puede recurrir a incrementar el tamaño de las muestras o, si esto no es práctico, se pueden usar métodos exactos de probabilidad que involucren la distribución multinomial.

COEFICIENTE DE CONTINGENCIA

La medida del grado de relación, asociación o dependencia de las clasificaciones en una tabla de contingencia está dada por

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \tag{26}$$

lo cual recibe el nombre de *coeficiente de contingencia*. Entre más grande sea el valor de C, mayor será el grado de asociación. El número de filas y columnas de la tabla de contingencia determina el valor máximo de C, el cual jamás es mayor que 1. Para una tabla $k \times k$ el valor máximo de C está dado por $\sqrt{(k-1)/k}$. *Véanse* problemas 7.52 y 7.53.

Problemas resueltos

PRUEBAS PARA MEDIAS Y PROPORCIONES USANDO DISTRIBUCIONES NORMALES

7.1. Encuentre la probabilidad de obtener entre 40 y 60 caras inclusive en 100 lanzamientos de una moneda balanceada.

Según la distribución binomial, la probabilidad requerida es

$${}_{100}C_{40}\left(\frac{1}{2}\right)^{40}\left(\frac{1}{2}\right)^{60} + {}_{100}C_{41}\left(\frac{1}{2}\right)^{41}\left(\frac{1}{2}\right)^{59} + \cdots + {}_{100}C_{60}\left(\frac{1}{2}\right)^{60}\left(\frac{1}{2}\right)^{40}$$

La media y la desviación estándar del número de caras en 100 lanzamientos están dadas por

$$\mu = np = 100\left(\frac{1}{2}\right) = 50 \qquad \sigma = \sqrt{npq} = \sqrt{(100)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)} = 5$$

Ya que np y nq son mayores que 5, la aproximación normal a la distribución binomial puede usarse para evaluar la suma anterior.

En una escala continua, entre 40 y 60 caras inclusive es lo mismo que entre 39.5 y 60.5 caras.

$$39.5 \text{ en unidades estándar} = \frac{39.5 - 50}{5} = -2.10 \qquad 60.5 \text{ en unidades estándar} = \frac{60.5 - 50}{5} = 2.10$$

$$\begin{aligned}\text{Probabilidad requerida} &= \text{área bajo la curva normal entre } z = -2.10 \text{ y } z = 2.10 \\ &= 2(\text{área entre } z = 0 \text{ y } z = 2.10) = 2(0.4821) = 0.9642\end{aligned}$$

7.2. Para probar la hipótesis de que una moneda es balanceada, se debe adoptar la siguiente regla de decisión: (*1*) aceptar la hipótesis si el número de caras en una muestra sencilla de 100 lanzamientos está entre 40 y 60 inclusive, (*2*) rechazar la hipótesis en otros casos.

(*a*) Encuentre la probabilidad de rechazar la hipótesis cuando en realidad es correcta.

(*b*) Interprete de manera gráfica la regla de decisión y el resultado en la parte (*a*).

(*c*) ¿Qué conclusiones saca usted si en la muestra de 100 lanzamientos obtiene 53 caras? ¿60 caras?

(*d*) ¿Podría usted estar equivocada en sus conclusiones con respecto a (*c*)? Explique.

(*a*) Según el problema 7.1, la probabilidad de no obtener entre 40 y 60 caras inclusive si la moneda es balanceada, es igual a $1 - 0.9642 = 0.0358$. Entonces, la probabilidad de rechazar la hipótesis cuando ésta es correcta es igual a 0.0358.

(*b*) La regla de decisión se ilustra en la figura 7-2, que muestra la probabilidad de la distribución de caras en 100 lanzamientos de una moneda balanceada.

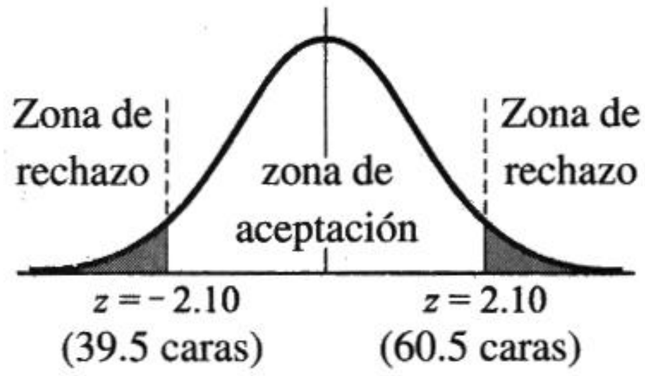


Fig. 7-2

Si de una muestra sencilla de 100 lanzamientos se obtiene un número de caras z entre -2.10 y 2.10, aceptamos la hipótesis; de otra manera, la rechazamos y concluimos que la moneda no es balanceada.

El error que se comete al rechazar la hipótesis cuando ésta se debe aceptar es el *error de tipo I* de la regla de decisión; y la probabilidad de cometer este error es igual a 0.0358 según la parte (*a*). Se representa por el área sombreada total de la figura.

Si en una muestra sencilla de 100 lanzamientos se obtiene un número z de caras, tal que esté ubicado en las regiones sombreadas, podemos decir que ese número z difiere *significativamente* de lo que esperamos si la hipótesis fuera verdadera. Por esta razón el área sombreada total (es decir, la probabilidad de error de tipo I) se denomina el *nivel de significancia* de la regla de decisión. Éste equivale a 0.0358 en este caso. Por lo tanto hablamos de rechazar la hipótesis al nivel de significancia de 0.0358 ó 3.58%.

(*c*) De acuerdo con la regla de decisión, en ambos casos tendremos que aceptar la hipótesis de que la moneda es balanceada. Uno podría argumentar que si sólo se hubieran obtenido una o más caras, habríamos podido rechazar la hipótesis. A esto nos enfrentamos cuando no hay una línea divisoria muy definida al tomar decisiones.

(*d*) Sí. Podemos aceptar la hipótesis cuando en realidad se debe rechazar, como sería el caso, por ejemplo, cuando la probabilidad de caras es 0.7 en lugar de 0.5.

El error cometido al aceptar la hipótesis cuando se debería rechazar es el *error de tipo II* de la decisión. Para más discusión al respecto *véanse* los problemas 7.23 a 7.25.

7.3. Diseñe una regla de decisión para probar la hipótesis de que una moneda es balanceada si se obtienen 64 lanzamientos y se utiliza un nivel de significancia de (*a*) 0.05, (*b*) 0.01.

(*a*) **Primer método.**

Si el nivel de significancia es 0.05, por simetría cada área sombreada de la figura 7-3 es 0.025. Así, el área entre 0 y z_1 es $0.5000 - 0.4750$ y $z_1 = 1.96$.

Entonces, una posible regla de decisión es:

(*1*) Aceptar la hipótesis de que la moneda es balanceada si Z está entre -1.96 y 1.96.

(*2*) De otra manera rechazar la hipótesis.

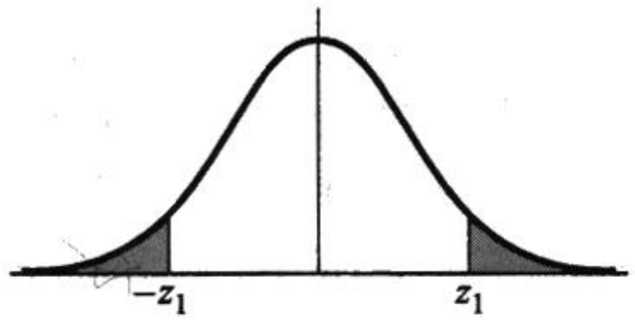


Fig. 7-3

Los valores críticos -1.96 y 1.96 también se pueden obtener de la tabla 7-1.

Para expresar esta decisión en términos del número de caras que se obtendrán en 64 lanzamientos de la moneda, observe que la media y la desviación estándar de la distribución binomial exacta están dadas por

$$\mu = np = 64(0.5) = 32 \qquad \text{y} \qquad \sigma = \sqrt{npq} = \sqrt{64(0.5)(0.5)} = 4$$

bajo la hipótesis de que la moneda es balanceada. Entonces, $Z = (X - \mu)/\sigma = (X - 32)/4$.

Si $Z = 1.96$, $(X - 32)/4 = 1.96$ o $X = 39.84$. Si $Z = -1.96$, $(X - 32)/4 = -1.96$ o $X = 24.16$. Por consiguiente, la regla de decisión es:

(*1*) Aceptar la hipótesis de que la moneda es balanceada si el número de caras está entre 24.16 y 39.85, es decir, entre 25 y 39 inclusive.

(*2*) De otra manera rechazar la hipótesis.

Segundo método.

Con probabilidad de 0.95, el número de caras estará entre $\mu - 1.96\sigma$ y $\mu + 1.96\sigma$; es decir, $np - 1.96\sqrt{npq}$ y $np + 1.96\sqrt{npq}$ o entre $32 - 1.96(4) = 24.16$ y $32 + 1.96(4) = 39.84$, lo cual lleva a la regla de decisión anterior.

Tercer método.

$-1.96 < Z < 1.96$ es equivalente a $-1.96 < (X - 32)/4 < 1.96$. Consecuentemente, $-1.96(4) < (X - 32) < 1.96(4)$ o $32 - 1.96(4) < X < 32 + 1.96(4)$, es decir, $24.16 < X < 39.84$, lo cual también conduce a la regla de decisión anterior.

(*b*) Si el nivel de significancia es 0.01, cada área sombreada en la figura 7-3 es 0.005. Entonces, el área entre 0 y z_1 es $0.5000 - 0.0050 = 0.4950$ y $z_1 = 2.58$ (más exactamente 2.575). Esto también se puede obtener de la tabla 7-1.

Siguiendo el procedimiento del segundo método de la parte (*a*), vemos que con probabilidad 0.99 el número de caras estará entre $\mu - 2.58\sigma$ y $\mu + 2.58\sigma$, es decir, $32 - 2.58(4) = 21.68$ y $32 + 2.58(4) = 42.32$.

Por consiguiente, la regla de decisión es:

(*1*) Aceptar la hipótesis si el número de caras está entre 22 y 42 inclusive.

(*2*) De otra manera rechazar la hipótesis.

7.4. ¿Cómo puede usted diseñar una regla de decisión en el problema 7.3 de manera que evite el error de tipo II?

Se comete un error de tipo II cuando se acepta una hipótesis que debe rechazarse. Para evitar este error, en lugar de aceptar la hipótesis, simplemente no la rechazamos, lo cual significa que en este caso nos abstenemos de tomar una decisión. Por ejemplo, podemos escribir la regla de decisión del problema 7.3(*b*) así:

(*1*) No rechazar la hipótesis si el número de caras está entre 22 y 42 inclusive.

(*2*) De otra manera rechazar la hipótesis.

Sin embargo, en muchos ejemplos prácticos es importante decidir si una hipótesis debe aceptarse o rechazarse. Una discusión completa de tales casos requiere consideraciones de error de tipo II (*véanse* los problemas 7.23 a 7.25).

7.5. En un experimento de percepción extrasensorial (EPE) se pide a un sujeto que está en una habitación que diga el color (rojo o azul) de una carta que selecciona otro sujeto, que está en otra habitación, de un naipe de 50 cartas bien barajadas. El sujeto desconoce el número de cartas rojas o azules que tiene el naipe. Si él identifica correctamente 32 cartas, determine si los resultados son significativos al nivel de significancia de (*a*) 0.05, (*b*) 0.01. (*c*) Encuentre e interprete el valor P de la prueba.

Si p es la probabilidad de que el sujeto identifique correctamente el color de la carta, entonces debemos decidir entre las siguientes hipótesis:

H_0: $p = 0.5$, y el sujeto sólo esta adivinando, es decir, los resultados están sujetos al azar.

H_1: $p > 0.5$, y el sujeto tiene poderes de percepción extrasensorial.

Escogemos la prueba de una cola ya que no estamos interesados en la habilidad de obtener valores extremadamente bajos sino en la habilidad de obtener valores altos.

Si la hipótesis H_0 es verdadera, la media y la desviación estándar del número de cartas identificadas correctamente están dadas por

$$\mu = np = 50(0.5) = 25 \qquad \text{y} \qquad \sigma = \sqrt{npq} = \sqrt{50(0.5)(0.5)} = \sqrt{12.5} = 3.54$$

(*a*) Para la prueba de una cola al nivel de significancia de 0.05, debemos escoger z_1 en la figura 7-4 de manera que el área sombreada en la región crítica de puntajes altos sea 0.05. Entonces el área entre 0 y z_1 es 0.4500 y $z_1 = 1.645$. Esto también se puede obtener de la tabla 7-1.

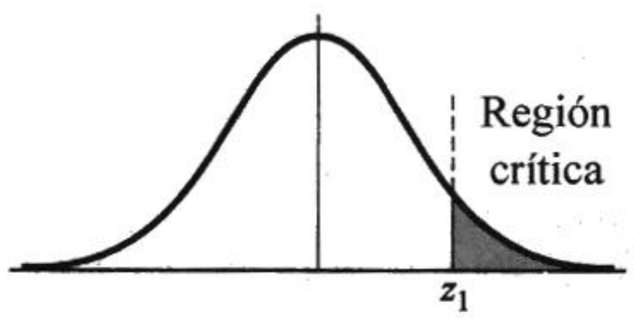


Fig. 7-4

Por consiguiente, nuestra regla de decisión o prueba de significancia es:

(*1*) Si el valor z observado es mayor que 1.645, los resultados son significativos al nivel de 0.05 y el sujeto tiene poderes extrasensoriales.

(*2*) Si el valor z es menor que 1.645, los resultados se deben al azar, es decir, no son significativos al nivel 0.05.

Dado que 32 en unidades estándar es $(32 - 25)/3.54 = 1.98$, lo cual es mayor que 1.645, la decisión (*1*) es válida, es decir, concluimos al nivel de 0.05 que el sujeto tiene poderes extrasensoriales.

Observe que realmente debemos aplicar una corrección por continuidad, ya que 32 en escala continua está entre 31.5 y 32.5. Sin embargo, 31.5 tiene un valor estándar de $(31.5 - 25)/3.54 = 1.84$, y se obtiene la misma conclusión.

(*b*) Si el nivel de significancia es 0.01, entonces el área entre 0 y z_1 es 0.4900 y $z_1 = 2.33$. Dado que 32 (o 31.50) en unidades estándar es 1.98 (o 1.84), lo cual es menor que 2.33, concluimos que los resultados *no son significativos* al nivel de 0.01.

Algunos estudiosos de la estadística adoptan la terminología de que los resultados significativos al nivel de 0.01 son *altamente significativos*, los resultados significativos al nivel 0.05, pero no al nivel 0.01, son *probablemente significativos*, mientras que los resultados significativos a niveles mayores que 0.05 son *no significativos*.

De acuerdo con esta terminología podemos concluir que los resultados del experimento anterior son *probablemente significativos*, así que se recomiendan investigaciones adicionales de dicho fenómeno.

(*c*) El valor P de la prueba es la probabilidad de que el color de 32 cartas o más se identifique correctamente en una selección al azar. El valor estándar de 32, teniendo en cuenta la corrección para continuidad, es $z = 1.84$. Por consiguiente, el valor P es $P(Z \geq 1.84) = 0.032$. Los estadísticos podrían decir que con base en el experimento, la probabilidad de concluir erróneamente que el sujeto tiene poderes extrasensoriales es de 3 en 100.

7.6. El fabricante de un medicamento patentado sostiene que tiene una eficacia del 90% en aliviar cierta alergia durante un periodo de 8 horas. En una muestra de 200 personas con dicha alergia, el medicamento mejoró a 160 personas. (*a*) Determine si la afirmación del fabricante es verídica usando un nivel de significancia de 0.01. (*b*) Encuentre el valor P para la prueba.

(*a*) Sea p la probabilidad de obtener alivio de la alergia al utilizar el medicamento. Entonces debemos decidir entre estas dos hipótesis:

H_0: $p = 0.9$ y la afirmación es correcta

H_1: $p < 0.9$ y la afirmación es falsa

Escogemos una prueba de una cola ya que estamos interesados en determinar si la proporción de personas aliviadas por el medicamento es muy baja.

Si el nivel de significancia se toma como 0.01, es decir, si el área sombreada de la figura 7-5 es 0.01, entonces, $z_1 = -2.33$ como se puede ver del problema 7.5(b), usando la simetría de la curva, o de la tabla 7-1.

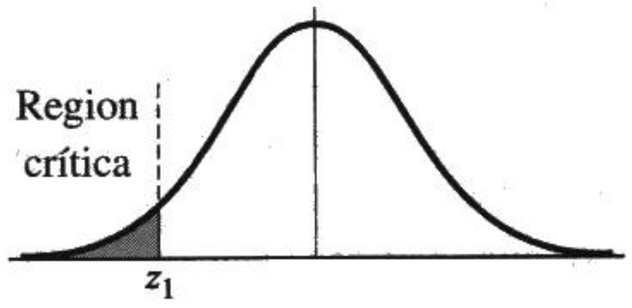


Fig. 7-5

Tomamos como nuestra regla de decisión:

(1) La afirmación no es verdadera si Z es menor que -2.33 (en cuyo caso rechazamos H_0).

(2) De otra manera, la afirmación es válida y los resultados observados se deben al azar (en cuyo caso aceptamos H_0).

Si H_0 es verdadera, $\mu = np = 200(0.9) = 180$ y $\sigma = \sqrt{npq} = \sqrt{(200)(0.9)(0.1)} = 4.23$.

Ahora 160 en unidades estándar es $(160 - 180)/4.23 = -4.73$, lo cual es mucho menos que -2.33. Así que según nuestra regla de decisión concluimos que la afirmación no es válida y que los resultados muestrales son *altamente significativos* (*véase* final del problema 7.5).

(b) El valor P de la prueba es $P(Z \leq -4.73) \approx 0$, lo cual muestra que la afirmación es casi completamente falsa. Eso es, si H_0 fuera cierta, sería casi seguro que una muestra aleatoria de 200 personas alérgicas que usaron el medicamento incluiría más de 160 personas que encontraron alivio.

7.7. Se calculó que el tiempo de vida medio de una muestra de 100 bombillas fluorescentes producidas por una compañía es de 1570 horas con desviación estándar de 120 horas. Si μ es el tiempo de vida medio de todas las bombillas producidas por la compañía, pruebe la hipótesis $\mu = 1600$ horas en contra de una hipótesis alterna $\mu \neq 1600$ horas, usando un nivel de significancia de (a) 0.05, (b) 0.01. (c) Encuentre el valor P de la prueba.

Debemos decidir entre dos hipótesis

$$H_0\colon\ \mu = 1600 \text{ horas y } H_1\colon\ \mu \neq 1600 \text{ horas}$$

Se debe usar una prueba de dos colas ya que $\mu \neq 1600$ incluye los valores más grandes y más pequeños que 1600.

(a) Para una prueba de dos colas con nivel de significancia de 0.05, tenemos la siguiente regla de decisión:

(1) Rechazar la H_0 si el valor z de la media muestral está por fuera del rango -1.96 a 1.96.

(2) De otra manera aceptar H_0 (o abstenerse de tomar una decisión).

El estadístico bajo consideración es la media muestral $\bar{X}$. La distribución muestral de $\bar{X}$ tiene media $\mu_{\bar{X}} = \mu$ y desviación estándar $\sigma_{\bar{X}} = \sigma/\sqrt{n}$, donde μ y σ son la media y la desviación estándar de la población de todas las bombillas producidas por la compañía.

Bajo la hipótesis H_0, tenemos $\mu = 1600$ y $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 120/\sqrt{100} = 12$, usando la desviación estándar muestral como un estimador de σ. Dado que $Z = (\bar{X} - 1600)/12 = (1570 - 1600)/12 = -2.50$ se encuentra fuera del rango de -1.96 a 1.96, rechazamos la H_0 al nivel de significancia de 0.05.

(b) Si el nivel de significancia es 0.01, el rango de -1.96 a 1.96 en la parte (a) de la decisión de la regla se remplaza por -2.58 a 2.58. Entonces, dado que el valor z de -2.50 está dentro de este rango, aceptamos H_0 (o nos abstenemos de tomar cualquier decisión) con un nivel de significancia de 0.01.

(c) El valor P de la prueba de dos colas es $P(Z \leq -2.50) + P(Z \geq 2.50) = 0.0124$, lo cual es la probabilidad de que el tiempo de vida media de menos de 1570 horas o más de 1630 horas ocurra por azar si H_0 fuera verdadera.

7.8. En el problema 7.7 pruebe la hipótesis $\mu = 1600$ horas contra la hipótesis alterna $\mu < 1600$ horas usando un nivel de significancia de (*a*) 0.05, (*b*) 0.01. (*c*). Encuentre el valor *P* de la prueba.

Debemos decidir entre dos hipótesis

$$H_0\text{: } \mu = 1600 \text{ horas y } H_1\text{: } \mu < 1600 \text{ horas}$$

Aquí se debe usar una prueba de una cola (*véase* figura 7-5).

(*a*) Si el nivel de significancia es 0.05, la región sombreada de la figura 7-5 tiene un área de 0.05, y encontramos que $z_1 = -1.645$. Por lo tanto, adoptamos la regla de decisión:

(*1*) Rechazar la H_0 si Z es menor que -1.645.

(*2*) De otra manera aceptar H_0 (o abstenerse de tomar cualquier decisión).

Dado que, tal como en el problema 7.7(*a*), el valor z es -2.50, lo cual es menor que -1.645, rechazamos la H_0 a un nivel de significancia de 0.05. Observe que esta decisión es idéntica a la que se llegó en el problema 7.7(*a*) usando la prueba de dos colas.

(*b*) Si el nivel de significancia es 0.01, el valor z_1 en la figura 7-5 es -2.33. Así, adoptamos nuestra regla de decisión:

(*1*) Rechazar H_0 si Z es menor que -2.33.

(*2*) De otra manera, aceptar H_0 (o abstenerse de tomar cualquier decisión).

Dado que, tal como en el problema 7.7(*a*), el valor z es -2.50, lo cual es menor que -2.33, rechazamos H_0 con un nivel de significancia de 0.01. Observe que esta decisión no es la misma a la que se llegó en el problema 7.7(*b*) usando la prueba de dos colas.

Se deduce que decisiones concernientes a hipótesis H_0 dadas con base en pruebas de una y dos colas no siempre concuerdan. Esto es de esperarse, por supuesto, ya que estamos probando H_0 en contra de una alternativa diferente en cada caso.

(*c*) El valor *P* de la prueba es $P(\bar{X} \leq 1570) = 0.0062$, que es la probabilidad de que ocurra un tiempo de vida medio de menos de 1570 horas al azar si H_0 es verdadera.

7.9. La resistencia al rompimiento de los cables producidos por un fabricante tienen media 1800 lb y desviación estándar 100 lb. Se afirma que con ayuda de una técnica nueva introducida en el proceso se puede aumentar dicha resistencia. Para probar esta afirmación, se probó una muestra de 50 cables, y se encontró que la resistencia promedio al rompimiento es 1850 lb. (*a*) ¿Podemos apoyar la afirmación con un nivel de significancia de 0.01? (*b*) ¿Cuál es el valor *P* de la prueba?

(*a*) Debemos decidir entre estas dos hipótesis:

H_0: $\mu = 1800$ lb, y realmente no hay cambio en la resistencia al rompimiento

H_1: $\mu > 1800$ lb, y hay cambio en la resistencia al rompimiento

Aquí se puede usar una prueba de una cola (*véase* figura 7-4). Con un nivel de significancia de 0.01 la regla de decisión es:

(*1*) Si el valor z observado es mayor que 2.33, los resultados son significativos al nivel de 0.01 y H_0 se rechaza.

(*2*) De otra manera se acepta H_0 (o se abstiene de tomar una decisión).

Bajo la hipótesis de que H_0 es verdadera, encontramos

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{1850 - 1800}{100/\sqrt{50}} = 3.55$$

que es mayor que 2.33. Entonces concluimos que los resultados son *altamente significativos* y debemos aceptar la afirmación.

(*b*) El valor *P* de la prueba es $P(Z \geq 3.55) = 0.0002$, lo cual es la probabilidad de que la resistencia al rompimiento de 1850 lb o más ocurra por azar si H_0 es verdadera.

PRUEBAS QUE INVOLUCRAN DIFERENCIAS DE MEDIAS Y PROPORCIONES

7.10. Se evaluaron dos cursos de 40 y 50 estudiantes, respectivamente. En el primer curso la nota promedio fue 74 con desviación estándar de 8, mientras que en el segundo curso la nota promedio fue 78 con desviación estándar de 7. ¿Existe alguna diferencia en el rendimiento de los dos cursos con nivel de significancia de (*a*) 0.05, (*b*) 0.01? ¿Cuál es el valor *P* de la prueba?

Supongamos que los dos cursos vienen de dos poblaciones con medias respectivas μ_1 y μ_2. Entonces tenemos que decidir entre estas hipótesis:

H_0: $\mu_1 = \mu_2$, y la diferencia se debe al azar

H_1: $\mu_1 \neq \mu_2$, y hay una diferencia significativa entre ambos cursos

Bajo la hipótesis H_0, ambos cursos vienen de la misma población. La media y la desviación estándar de la diferencia de medias está dada por

$$\mu_{\bar{X}_1-\bar{X}_2} = 0 \qquad \sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} = \sqrt{\frac{8^2}{40}+\frac{7^2}{50}} = 1.606$$

donde usamos las desviaciones estándar muestrales como estimadores de σ_1 y σ_2.

Entonces, $$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1-\bar{X}_2}} = \frac{74-78}{1.606} = -2.49$$

(*a*) Para una prueba de dos colas, los resultados son significativos al nivel de 0.05 si *Z* está por fuera del rango −1.96 a 1.96. Entonces concluimos que al nivel de 0.05 hay una diferencia significativa en el desempeño de los dos cursos y que probablemente el segundo curso es mejor.

(*b*) Para una prueba de dos colas los resultados son significativos al nivel de 0.01 si *Z* está por fuera del rango −2.58 y 2.58. Entonces concluimos que al nivel de 0.01 no hay diferencia significativa entre los dos cursos.

Dado que los resultados son significativos al nivel de 0.05, pero no lo son al nivel de 0.01, concluimos que los resultados son *probablemente significativos*, de acuerdo con la terminología usada al final del problema 7.5.

(*c*) El valor *P* de la prueba de dos colas es $P(Z \leq -2.49) + P(Z \geq 2.49) = 0.0128$, que es la probabilidad de que los estadísticos observados ocurran en la misma población.

7.11. La estatura promedio de 50 estudiantes varones que mostraron una participación por encima del promedio en las competencias escolares, fue 68.2 pulgadas con desviación estándar de 2.5 pulgadas, mientras que 50 estudiantes varones que no mostraron interés en tal participación tuvieron estatura promedio de 67.5 pulgadas con desviación estándar de 2.8 pulgadas. (*a*) Pruebe la hipótesis de que los estudiantes varones que participaron en las pruebas atléticas son más altos que los demás estudiantes varones. (*b*) ¿Cuál es el valor *P* de la prueba?

(*a*) Debemos decidir entre las hipótesis

H_0: $\mu_1 = \mu_2$, y no hay diferencia entre las estaturas promedio

H_1: $\mu_1 > \mu_2$, y la estatura promedio del primer grupo es mayor que la del segundo grupo

Bajo la hipótesis H_0,

$$\mu_{\bar{X}_1-\bar{X}_2} = 0 \qquad \sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(2.5)^2}{50}+\frac{(2.8)^2}{50}} = 0.53$$

donde usamos las desviaciones estándar muestrales como estimadores de σ_1 y σ_2.

Entonces $$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1-\bar{X}_2}} = \frac{68.2-67.5}{0.53} = 1.32$$

Sobre la base de una prueba de una cola y un nivel de significancia de 0.05, podemos rechazar la hipótesis H_0 si el valor z fuera mayor que 1.645. Por consiguiente, no podemos rechazar la hipótesis a este nivel de significancia.

Debe notarse, sin embargo, que la hipótesis se puede rechazar al nivel de 0.10 si queremos tomar el riesgo de equivocarnos con probabilidad de 0.10, es decir, con probabilidad de 1 en 10.

(*b*) El valor P de la prueba es $P(Z \geq 1.32) = 0.0934$, que es la probabilidad de que la diferencia positiva observada entre las estaturas promedio de los atletas masculinos y otros estudiantes varones, ocurra al azar si H_0 fuera verdadera.

7.12. ¿En cuánto se debe aumentar el tamaño de la muestra de cada uno de los dos grupos en el problema 7.11 para que la diferencia observada de 0.7 pulgadas en las estaturas medias sea significativa al nivel de significancia de (*a*) 0.05, (*b*) 0.01?

Supongamos que el tamaño de la muestra de cada grupo es n y que las desviaciones estándar de los dos grupos son las mismas. Entonces, bajo las mismas hipótesis H_0 tenemos $\mu_{\bar{X}_1 - \bar{X}_2} = 0$ y

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}} = \sqrt{\frac{(2.5)^2 + (2.8)^2}{n}} = \sqrt{\frac{14.09}{n}} = \frac{3.75}{\sqrt{n}}$$

para una diferencia observada en las estaturas promedio de 0.7 pulgadas,

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}} = \frac{0.7}{3.75/\sqrt{n}} = \frac{0.7\sqrt{n}}{3.75}$$

(*a*) La diferencia observada será significativa a un nivel de 0.05 si

$$\frac{0.7\sqrt{n}}{3.75} \geqq 1.645 \quad \text{o} \quad \sqrt{n} \geqq 8.8 \quad \text{o} \quad n \geqq 78$$

Por consiguiente, debemos aumentar el tamaño de la muestra de cada grupo, por lo menos en $78 - 50 = 28$.

(*b*) La diferencia observada será significativa a un nivel de 0.01 si

$$\frac{0.7\sqrt{n}}{3.75} \geqq 2.33 \quad \text{o} \quad \sqrt{n} \geqq 12.5 \quad \text{o} \quad n \geqq 157$$

Entonces, debemos incrementar el tamaño de la muestra en cada grupo por lo menos en $157 - 50 = 107$.

7.13. Tenemos dos grupos, A y B, cada uno de 100 personas que sufren una enfermedad. Se le da un suero al grupo A, pero no al grupo B (*grupo de control*); para lo demás, ambos grupos se manejan de manera idéntica. Se encontró que en los grupos A y B, 75 y 65 personas, respectivamente, se recuperaron de la enfermedad. Pruebe la hipótesis de que el suero ayuda a curar la enfermedad usando un nivel de significancia de (*a*) 0.01, (*b*) 0.05, (*c*) 0.10. (*d*) Encuentre el valor P de la prueba.

Sean p_1 y p_2, respectivamente, las proporciones de la población curada (*1*) usando el suero, (*2*) sin usar el suero. Debemos decidir entre estas dos hipótesis.

H_0: $p_1 = p_2$, y las diferencias observadas se deben al azar, es decir, el suero no es efectivo

H_1: $p_1 > p_2$ y el suero es efectivo

Bajo la hipótesis H_0,

$$\mu_{P_1 - P_2} = 0 \quad ; \quad \sigma_{P_1 - P_2} = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{(0.70)(0.30)\left(\frac{1}{100} + \frac{1}{100}\right)} = 0.0648$$

donde usamos como estimado de p la proporción promedio de curas en los dos grupos muestrales, dados por $(75 + 65)/200 = 0.70$, y donde $q = 1 - p = 0.30$. Entonces

$$Z = \frac{P_1 - P_2}{\sigma_{P_1 - P_2}} = \frac{0.750 - 0.650}{0.0648} = 1.54$$

(a) Con base en una prueba de una cola con nivel de significancia de 0.01, podemos rechazar la hipótesis H_0 sólo si el valor z fuera mayor que 2.33. Dado que el valor z es sólo 1.54, debemos concluir que, con este nivel de significancia, los resultados se deben al azar.

(b) Con base en una prueba de una cola con nivel de significancia de 0.05, podemos rechazar la hipótesis H_0 sólo si el valor z fuera mayor que 1.645. Debemos concluir que, con este nivel de significancia, los resultados también se deben al azar.

(c) Si se hubiera usado una prueba de una cola con nivel de significancia de 0.10, podríamos rechazar la hipótesis H_0 sólo si el valor z fuera mayor que 1.28. Dado que esta condición se satisface, podemos concluir que el suero es efectivo con un nivel de significancia de 0.10.

(d) El valor P de la prueba es $P(Z \geq 1.54) = 0.0618$, que es la probabilidad de que un valor z de 1.54 o mayor, a favor del grupo que usó el suero, ocurra por azar si H_0 es verdadera.

Observe que nuestras conclusiones anteriores dependen de cuánto estemos dispuestos a arriesgar al equivocarnos. Si en realidad los resultados se deben al azar y nosotros concluimos que se deben al suero (error de tipo I), debemos proceder a dar el medicamento a grupos más numerosos de personas, sólo para encontrar que éste no es efectivo. Y no siempre estamos dispuestos a asumir este riesgo.

Por otro lado, podemos concluir que el suero no ayuda, cuando sí lo hace (error tipo II). Tal conclusión es muy peligrosa, especialmente si hay vidas humanas de por medio.

7.14. Desarrolle el problema 7.13 si cada grupo es de 300 personas, y 225 del grupo A y 195 del grupo B se curan.

En este caso la proporción de gente curada en los dos grupos es, respectivamente, $225/300 = 0.750$ y $195/300 = 0.650$, que es lo mismo que en el problema 7.13. Bajo la hipótesis H_0,

$$\mu_{P_1 - P_2} = 0 \quad ; \quad \sigma_{P_1 - P_2} = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{(0.70)(0.30)\left(\frac{1}{300} + \frac{1}{300}\right)} = 0.0374$$

donde $(225 + 195)/600 = 0.70$ se usa como estimado de p. Entonces

$$Z = \frac{P_1 - P_2}{\sigma_{P_1 - P_2}} = \frac{0.750 - 0.650}{0.0374} = 2.67$$

Dado que el valor de z es mayor que 2.33, podemos rechazar la hipótesis con un nivel de significancia de 0.01, es decir, concluimos que el suero no es efectivo sólo con una probabilidad de 0.01 de estar equivocados. Aquí el valor P de la prueba es $P(Z \geq 2.67) = 0.0038$.

Esto demuestra que aumentar el tamaño de la muestra puede incrementar la confiabilidad de las decisiones. Sin embargo, en muchos casos, no es práctico aumentar el tamaño de la muestra, y estamos obligados a tomar decisiones con base en la información disponible y a enfrentarnos al mayor riesgo de tomar decisiones equivocadas.

7.15. Una encuesta hecha a 300 votantes del distrito A y a 200 del distrito B mostró que 56% y 48%, respectivamente, estaban a favor de un candidato dado. Con un nivel de significancia de 0.05 pruebe la hipótesis de que (a) hay diferencia entre los distritos, (b) el candidato es preferido en el distrito A. (c) Encuentre los valores P respectivos de las pruebas.

Sean p_1 y p_2 las proporciones de todos los votantes de los distritos A y B, respectivamente, que están a favor del candidato.

Bajo la hipótesis H_0: $p_1 = p_2$, tenemos

$$\mu_{P_1-P_2} = 0 \qquad \sigma_{P_1-P_2} = \sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \sqrt{(0.528)(0.472)\left(\frac{1}{300}+\frac{1}{200}\right)} = 0.0456$$

donde usamos como estimadores de p y q los valores $[(0.56)(300) + (0.48)(200)]/500 = 0.528$ y $1 - 0.528 = 0.472$, tenemos

$$Z = \frac{P_1 - P_2}{\sigma_{P_1-P_2}} = \frac{0.560 - 0.480}{0.0456} = 1.75$$

(*a*) Si solamente queremos determinar si hay una diferencia entre los distritos, debemos decidir entre la hipótesis H_0: $p_1 = p_2$ y H_1: $p_1 \neq p_2$, lo que involucra una prueba de dos colas.

Con base en una prueba de dos colas con nivel de significancia de 0.05, debemos rechazar H_0 si Z está fuera del intervalo -1.96 a 1.96. Dado que $Z = 1.75$ está dentro del intervalo, no podemos rechazar H_0 con este nivel de significancia, es decir, no hay diferencia significativa entre los dos distritos.

(*b*) Si queremos determinar si el candidato es preferido en el distrito A, debemos decidir entre las hipótesis H_0: $p_1 = p_2$ y H_1: $p_1 > p_2$, lo que involucra una prueba de una cola.

Con base en una prueba de una cola al nivel de significancia de 0.05, rechazamos H_0 si Z es mayor que 1,645. Dado que éste es el caso, rechazamos H_0 a este nivel y concluimos que el candidato es preferido en el distrito A.

(*c*) En la parte (*a*), el valor P es $P(Z \leq -1.75) + P(Z \geq 1.75) = 0.0802$, y el valor P en la parte (*b*) es $P(Z \geq 1.75) = 0.0401$.

PRUEBAS QUE INVOLUCRAN LA DISTRIBUCIÓN *t* DE STUDENT

7.16. En el pasado una máquina produjo empaques cuyo grosor medio era 0.050 cm. Para determinar si la máquina estaba trabajando correctamente se tomó una muestra de 10 empaques, para la cual el grosor promedio era 0.053 cm y su desviación estándar era 0.003 cm. Pruebe la hipótesis de que la máquina estaba trabajando adecuadamente usando un nivel de significancia de (*a*) 0.05, (*b*) 0.01. (*c*) Encuentre el valor P de la prueba.

Queremos decidir entre las hipótesis

H_0: $\mu = 0.050$, y la máquina estaba trabajando adecuadamente

H_1: $\mu \neq 0.050$, y la máquina no estaba trabajando adecuadamente

así que se requiere una prueba de dos colas.

Bajo la hipótesis H_0 tenemos $T = \dfrac{\bar{X} - \mu}{S}\sqrt{n-1} = \dfrac{0.053 - 0.050}{0.003}\sqrt{10-1} = 3.00$.

(*a*) Para una prueba de dos colas con nivel de significancia de 0.05, adoptamos la regla de decisión:

(*1*) Aceptar H_0 si T está en el intervalo de $-t_{0.975}$ a $t_{0.975}$, que es el intervalo de -2.26 a 2.26 para $10 - 1 = 9$ grados de libertad.

(*2*) De otra manera rechazar H_0.

Dado que $T = 3.00$, rechazamos H_0 al nivel de 0.05.

(*b*) Para una prueba de dos colas al nivel de significancia de 0.01, adoptamos la regla de decisión:

(*1*) Aceptar H_0 si T está dentro del intervalo de $-t_{0.995}$ a $t_{0.995}$, que es el intervalo de -3.25 a 3.25 para $10 - 1 = 9$ grados de libertad.

(*2*) De otra manera rechazar H_0.

Dado que $T = 3.00$, aceptamos la H_0 al nivel de 0.01.

Puesto que podemos rechazar H_0 al nivel de 0.05, pero no al nivel de 0.01, podemos decir que el resultado de la muestra es probable significativamente (*véase* la terminología al final del problema 7.5). Sería recomendable, por consiguiente, revisar la máquina o al menos tomar otra muestra.

(*c*) El valor P es $P(T \geq 3) + P(T \leq 3)$. La tabla del apéndice D muestra que $0.01 < P < 0.02$. Usando programas de computador encontramos que $P = 0.015$.

7.17. Una prueba de resistencia al rompimiento de 6 cuerdas fabricadas por una compañía mostró una resistencia promedio de 7750 lb con desviación estándar de 145 lb, mientras que el fabricante reclama que ésta es de 8000 lb. ¿Podemos apoyar la queja del fabricante con un nivel de significancia del (*a*) 0.05, (*b*) 0.01? ¿Cuál es el valor P de la prueba?

Debemos decidir entre las hipótesis

H_0: $\mu = 8000$ lb, y la queja del fabricante es justificada

H_1: $\mu < 8000$ lb, y la queja del fabricante no es justificada

así que se requiere una prueba de una cola.

Bajo la hipótesis H_0 tenemos

$$T = \frac{\bar{X} - \mu}{S}\sqrt{n-1} = \frac{7750 - 8000}{145}\sqrt{6-1} = -3.86.$$

(*a*) Para una prueba de una cola con nivel de significancia de 0.05, adoptamos la regla de decisión:

(*1*) Aceptar H_0 si T es mayor que $-t_{0.95}$, que para $6 - 1 = 5$ grados de libertad significa $T > -2.01$.

(*2*) De otra manera rechazar H_0.

Dado que $T = -3.86$, rechazamos H_0.

(*b*) Para una prueba de una cola con nivel de significancia de 0.01, adoptamos la regla de decisión:

(*1*) Aceptar H_0 si T es mayor que $-t_{0.99}$, que para 5 grados de libertad significa $T > -3.36$.

(*2*) De otra manera rechazar H_0.

Dado que $T = -3.86$, rechazamos H_0.

Concluimos que es extremadamente improbable que la queja del fabricante sea justificada.

(*c*) El valor P es $P(T \leq -3.86)$. La tabla en el apéndice D muestra $0.005 < P < 0.01$. Usando programas de computador, $P = 0.006$.

7.18. El CI (coeficiente de inteligencia) de 16 estudiantes de un área de la ciudad mostró una media de 107 con desviación estándar de 10, mientras que el CI de 14 estudiantes de otra área de la ciudad mostró una media de 112 con desviación estándar de 8. ¿Hay alguna diferencia significativa entre el CI de los dos grupos con un nivel de significancia de (*a*) 0.01, (*b*) 0.05? (*c*) ¿Cuál es el valor P de la prueba?

Si μ_1 y μ_2 son las medias del CI de la población de estudiantes de las dos áreas, tenemos que decidir entre las hipótesis

H_0: $\mu_1 = \mu_2$, y básicamente no hay diferencia entre los dos grupos

H_1: $\mu_1 \neq \mu_2$, y hay una diferencia significativa entre los dos grupos

Bajo la hipótesis H_0 tenemos

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{1/n_1 + 1/n_2}} \qquad \text{donde} \qquad \sigma = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}}$$

Entonces

$$\sigma = \sqrt{\frac{16(10)^2 + 14(8)^2}{16 + 14 - 2}} = 9.44 \qquad \text{y} \qquad T = \frac{112 - 107}{9.44\sqrt{1/16 + 1/14}} = 1.45$$

(*a*) Con base en una prueba de dos colas con nivel de significancia de 0.01, podemos rechazar la H_0 si T está fuera del rango de $-t_{0.995}$ a $t_{0.995}$, que, para $n_1 + n_2 - 2 = 16 + 14 - 2 = 28$ grados de libertad, es el rango de -2.76 a 2.76.

Por consiguiente, no podemos rechazar H_0 al nivel de significancia de 0.01.

(*b*) Con base en una prueba de dos colas con nivel de significancia de 0.05, podemos rechazar la hipótesis H_0 si T está fuera del rango $-t_{0.975}$ a $t_{0.975}$, que para 28 grados de libertad es el rango -2.05 a 2.05.

Por consiguiente, no podemos rechazar la hipótesis H_0 al nivel de significancia del 0.05. Concluimos que no hay diferencia significativa entre los CI de los dos grupos.

(*c*) El valor P de la prueba es $P(T \geq 1.45) + P(T \leq -1.45)$. La tabla del apéndice D muestra $0.1 < P < 0.2$. Por computador, $P = 0.158$.

7.19. En una estación agrícola se quiere probar el efecto de cierto fertilizante sobre la producción de trigo. Para lograrlo se seleccionan 24 parcelas de tierra de áreas iguales. La mitad se trata con fertilizante y la otra mitad no (grupo de control); para lo demás, las condiciones permanecen iguales. La producción promedio de trigo de las parcelas sin tratamiento fue 4.8 toneladas con desviación estándar de 0.40 toneladas, mientras que la producción media de las parcelas tratadas fue 5.1 toneladas con desviación estándar de 0.36 toneladas. ¿Podemos concluir que hay una mejora significativa en la producción de trigo debida al fertilizante si se usan niveles de significancia de (*a*) 1%, (*b*) 5%? (*c*)¿Cuál es el valor P de la prueba?

Si μ_1 y μ_2 son la producción media de trigo en tierra con y sin tratamiento, respectivamente, debemos decidir entre las hipótesis

H_0: $\mu_1 = \mu_2$, y la diferencia se debe al azar

H_1: $\mu_1 > \mu_2$, y el fertilizante mejora la producción

Bajo la hipótesis H_0 tenemos

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{1/n_1 + 1/n_2}} \qquad \text{donde} \qquad \sigma = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}}$$

Entonces

$$\sigma = \sqrt{\frac{12(0.40)^2 + 12(0.36)^2}{\text{i2} + 12 - 2}} = 0.397 \qquad \text{y} \qquad T = \frac{5.1 - 4.8}{0.397\sqrt{1/12 + 1/12}} = 1.85$$

(*a*) Con base en una prueba de una cola al nivel de significancia de 0.01, podemos rechazar la hipótesis si T es mayor que $t_{0.99}$, que, para $n_1 + n_2 - 2 = 12 + 12 - 2 = 22$ grados de libertad, es 2.51.

Por lo tanto, no podemos rechazar la hipótesis H_0 al nivel de significancia de 0.01.

(*b*) Con base en una prueba de una cola al nivel de significancia de 0.05, podemos rechazar la hipótesis si T es mayor que $t_{0.95}$, que 22 grados de libertad es 1.72.

Por lo tanto, podemos rechazar la hipótesis H_0 al nivel de significancia de 0.05.

Concluimos que la mejora en la producción de trigo por el uso del fertilizante es probablemente significativa. Sin embargo, sería deseable tener más evidencia antes de establecer conclusiones definitivas con relación a la inutilidad del fertilizante.

(*c*) El valor P para la prueba es $P(T \geq 1.85)$. La tabla del apéndice D muestra $0.025 < P < 0.05$. Haciendo los cálculos en computador, $P = 0.039$.

PRUEBAS QUE INVOLUCRAN LA DISTRIBUCIÓN CHI CUADRADO

7.20. En el pasado la desviación estándar del peso de ciertos paquetes de 40 onzas llenados por una máquina era 0.25 onzas. Una muestra aleatoria de 20 paquetes mostró una desviación estándar de 0.32 onzas. ¿Es el aumento aparente de la variación significativo al nivel de (*a*) 0.05, (*b*) 0.01? ¿Cuál es el valor P de la prueba?

Debemos escoger entre las hipótesis

H_0: $\sigma = 0.25$ onzas, y los resultados observados obedecen al azar

H_1: $\sigma > 0.25$ onzas, y la variación se incrementó

El valor de χ^2 para la muestra es $\chi^2 = ns^2/\sigma^2 = 20(0.32)^2/(0.25)^2 = 32.8$.

(*a*) Usando una prueba de una cola podemos rechazar la hipótesis H_0 al nivel de significancia de 0.05 si el valor de la muestra de χ^2 es mayor que $\chi^2_{0.95}$, el cual es igual a 30.1 para $\nu = 20 - 1 = 19$ grados de libertad. Por consiguiente, rechazamos la hipótesis H_0 al nivel de significancia de 0.05.

(*b*) Usando una prueba de una cola podemos rechazar la hipótesis H_0 al nivel de significancia de 0.01 si el valor de la muestra de χ^2 es mayor que $\chi^2_{0.99}$, el cual es igual a 36.2 para 19 grados de libertad. Por consiguiente, rechazamos la hipótesis H_0 al nivel de significancia de 0.01.

Concluimos que la variación probablemente se incrementó. Se debe examinar la máquina.

(*c*) El valor P es $P(\chi^2 \geq 32.8)$. La tabla del apéndice D muestra $0.025 < P < 0.05$. Por computador, $P = 0.0253$.

PRUEBAS QUE INVOLUCRAN LA DISTRIBUCIÓN *F*

7.21. Un instructor tiene dos cursos, *A* y *B*, en una materia en particular. El curso *A* tiene 16 estudiantes mientras que el *B* tiene 25. En el mismo examen, a pesar de que no hubo diferencias significativas en las notas promedio, el curso *A* tuvo una desviación estándar de 9 mientras que el *B* tuvo una desviación estándar de 12. ¿Podemos concluir que al nivel de significancia de (*a*) 0.01, (*b*) 0.05, la variabilidad del curso *B* es mayor que el de *A*? (*c*) ¿Cuál es el valor *P* de la prueba?

(*a*) Usemos los subíndices 1 y 2 para lols cursos *A* y *B*, respectivamente. Tenemos entonces, $s_1 = 9$, $s_2 = 12$, de manera que

$$\hat{s}_1^2 = \frac{n_1}{n_1 - 1} s_1^2 = \frac{16}{15}(9)^2 = 86.4, \qquad \hat{s}_2^2 = \frac{n_2}{n_2 - 1} s_2^2 = \frac{25}{24}(12)^2 = 150$$

Debemos decidir entre las hipótesis

H_0: $\sigma_1 = \sigma_2$, y cualquier variación se debe al azar

H_1: $\sigma_2 > \sigma_1$, y la variación del curso *B* es mayor que la del *A*

Por lo tanto, la decisión se debe basar en una prueba de una cola de la distribución *F*. Para las muestras en cuestión,

$$F = \frac{\hat{s}_2^2}{\hat{s}_1^2} = \frac{150}{86.4} = 1.74$$

El número de grados de libertad asociados con el numerador es $r_2 = 25 - 1 = 24$; para el denominador, $r_1 = 16 - 1 = 15$ grados de libertad. Al nivel de 0.01 para 24, 15 grados de libertad tenemos del apéndice F, $F_{0.99} = 3.99$. Entonces, dado que $F < F_{0.99}$, no podemos rechazar H_0 al nivel de 0.01.

(*b*) Dado que $F_{0.95} = 2.29$ para 24, 15 grados de libertad (*véase* apéndice F), vemos que $F < F_{0.95}$. Es así que tampoco podemos rechazar la H_0 al nivel de 0.05.

(*c*) El valor *P* de la prueba es $P(F \geq 1.74)$. Las tablas del apéndice F muestran que $F > F_{0.05}$. En computador, $P = 0.134$.

7.22. En el problema 7.21, ¿cambiarían sus conclusiones si resulta que hubo una diferencia significativa en las notas promedio de los cursos? Explique su respuesta.

Ya que las medias observadas de las notas no se usaron en el problema 7.21, no importa cuáles sean sus valores. Esto es de esperarse por cuanto no estamos tratando de decidir si hay una diferencia en el promedio de las notas, sino si hay una diferencia en la variación de éstas.

CURVAS DE OPERACIÓN CARACTERÍSTICAS

7.23. En relación con el problema 7.2, ¿cuál es la probabilidad de aceptar la hipótesis de que la moneda es balanceada cuando la probabilidad actual de caras es $p = 0.7$?

La hipótesis H_0 de que la moneda es balanceada, es decir, $p = 0.5$, se acepta cuando el número de caras en 100 lanzamientos se encuentra entre 39.5 y 60.5. La probabilidad de rechazar H_0 cuando deberíamos aceptarla (es decir, la probabilidad de error de tipo I) se representa por el área total α de la región sombreada bajo la curva normal a la izquierda en la figura 7-6. Como se calculó en el problema 7.2(a), esta área α, que representa el nivel de significancia de la prueba de H_0, es igual a 0.0358.

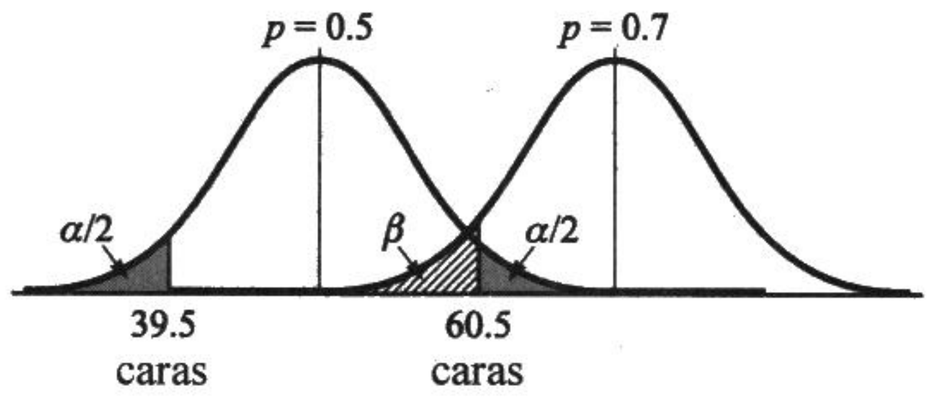


Fig. 7-6

Si la probabilidad de caras es $p = 0.7$, entonces la distribución de caras en 100 lanzamientos se representa por la curva normal a la derecha de la figura 7-6. La probabilidad de aceptar H_0 cuando en realidad $p = 0.7$ (o sea, la probabilidad de error tipo II) está dada por el área rayada β. Para calcular esta área observamos que la distribución bajo la hipótesis $p = 0.7$ tiene media y distribución estándar dadas por

$$\mu = np = (100)(0.7) = 70 \qquad \sigma = \sqrt{npq} = \sqrt{(100)(0.7)(0.3)} = 4.58$$

$$60.5 \text{ en unidades estándar} = \frac{60.5 - 70}{4.58} = -2.07$$

$$39.5 \text{ en unidades estándar} = \frac{39.5 - 70}{4.58} = -6.66$$

Entonces, β = área bajo la curva normal entre $z = -6.66$ y $z = -2.07 = 0.0192$.

Por consiguiente, con la regla de decisión dada, hay poca probabilidad de aceptar la hipótesis de que la moneda es balanceada cuando en realidad $p = 0.7$.

Observe que en este problema nos dieron la regla de decisión a partir de la cual calculamos α y β. En la práctica pueden surgir otras dos posibilidades.

(*1*) Decidimos sobre α (tal como 0.05 ó 0.01), llegamos a una regla de decisión, y luego calculamos β.

(*2*) Decidimos sobre α y β y luego llegamos a una regla de decisión.

7.24. Desarrolle el problema 7.23 si (a) $p = 0.6$, (b) $p = 0.8$, (c) $p = 0.9$, (d) $p = 0.4$.

(a) Si $p = 0.6$, la distribución de caras tiene media y desviación estándar dadas por

$$\mu = np = (100)(0.6) = 60 \qquad \sigma = \sqrt{npq} = \sqrt{(100)(0.6)(0.4)} = 4.90$$

$$60.5 \text{ en unidades estándar} = \frac{60.5 - 60}{4.90} = 0.102$$

$$39.5 \text{ en unidades estándar} = \frac{39.5 - 60}{4.90} = -4.18$$

Entonces, β = área bajo la curva normal estándar entre $z = -4.18$ y $z = 0.102 = 0.5405$.

Por consiguiente, con la regla de decisión dada hay bastante probabilidad de aceptar la hipótesis de que la moneda es balanceada cuando $p = 0.6$.

(*b*) Si $p = 0.8$, entonces $\mu = np = (100)(0.8) = 80$ y $\sigma = \sqrt{npq} = \sqrt{(100)(0.08)(0.2)} = 4$.

$$60.5 \text{ en unidades estándar} = \frac{60.5 - 80}{4} = -4.88$$

$$39.5 \text{ en unidades estándar} = \frac{39.5 - 80}{4} = -10.12$$

Entonces, β = área bajo la curva normal estándar entre $z = -10.12$ y $z = -4.88 = 0.0000$, muy cercano.

(*c*) A partir de la comparación con (*b*) o calculando, vemos que si $p = 0.9$, $\beta = 0$ para todo propósito práctico.

(*d*) Por simetría, $p = 0.4$ se obtiene el mismo valor de β como en $p = 0.6$, o sea, $\beta = 0.5405$.

7.25. Represente los resultados de los problemas 7.23 y 7.24 en una gráfica cuyos ejes sean (*a*) β y p, (*b*) $(1 - \beta)$ y p. Interprete las gráficas obtenidas.

La tabla 7-3 muestra los valores β correspondientes a los valores dados de p, como se obtuvieron en los problemas 7.23 y 7.24.

Observe que β representa la probabilidad de aceptar la hipótesis $p = 0.5$ cuando en realidad p es un valor diferente de 0.5. Sin embargo, si en realidad es verdad que $p = 0.5$, podemos interpretar β como la probabilidad de aceptar $p = 0.5$ cuando ésta se debe aceptar. Esta probabilidad es igual a $1 - 0.0358 = 0.9642$, y se introdujo en la tabla 7-3.

Tabla 7-3

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
β	0.0000	0.0000	0.0192	0.5405	0.9642	0.5405	0.0192	0.0000	0.0000

(*a*) La gráfica de β contra p, que aparece en la figura 7-7(*a*), se llama *curva de operación característica*, o *curva OC*, de la regla de decisión o prueba de hipótesis.

La distancia desde el punto máximo de la curva OC a la línea $\beta = 1$ es igual a $\alpha = 0.0358$, el nivel de significancia de la prueba.

En general, entre más empinada sea la curva OC, mejor será la regla de decisión para rechazar la hipótesis de que no es válida.

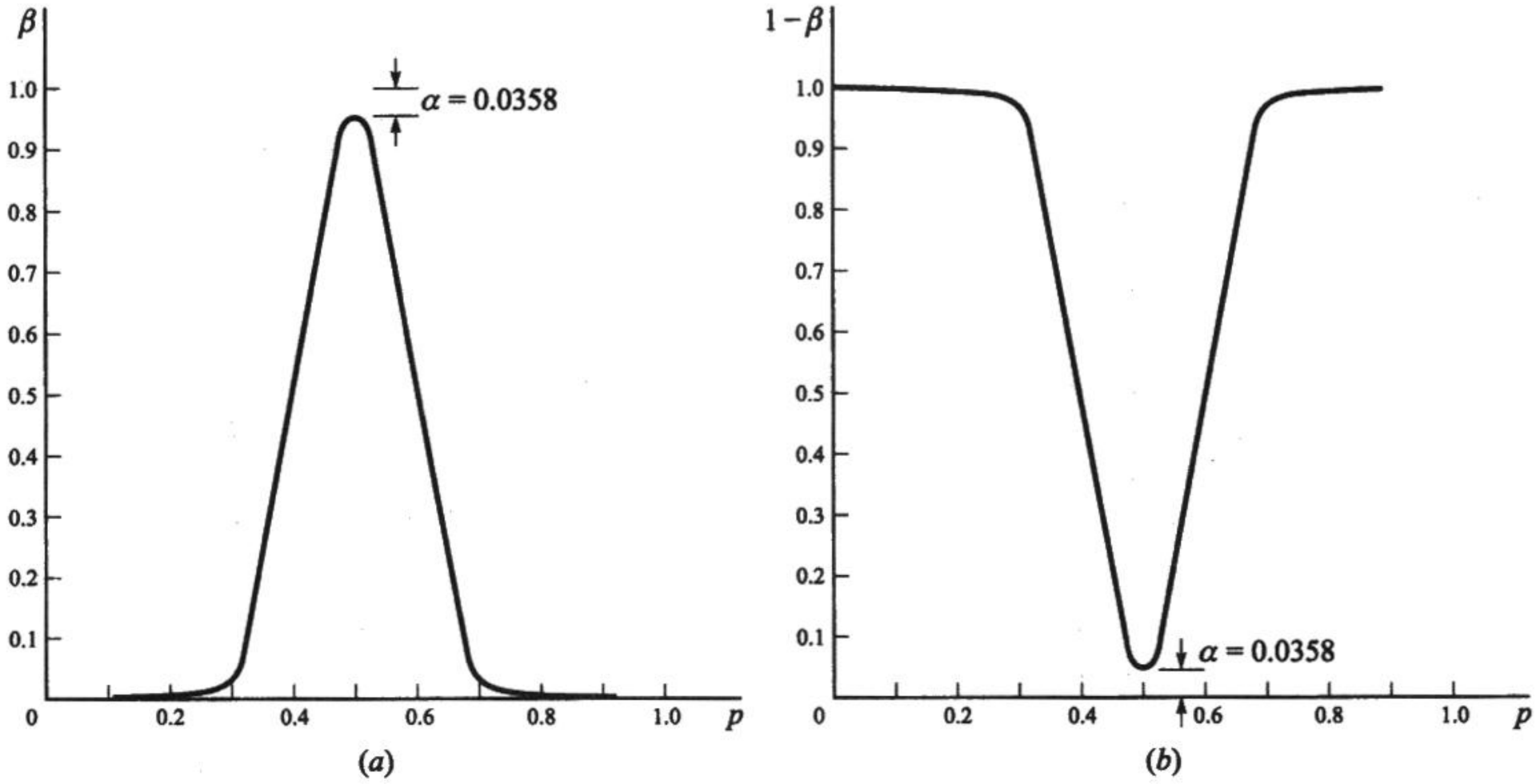


Fig. 7-7

(*b*) La gráfica de $(1 - \beta)$ contra p, que se muestra en la figura 7-7(*b*), se llama *curva de potencia* de la regla de decisión o prueba de hipótesis. Esta curva se obtiene simplemente invirtiendo la curva OC, de manera que en realidad ambas gráficas son equivalentes.

La cantidad $1 - \beta$ se denomina *función de potencia* ya que indica la habilidad o potencia de una prueba para rechazar las hipótesis que son falsas, es decir, que se deben rechazar. La cantidad β también se denomina *función de operación característica* de una prueba.

7.26. Una compañía fabrica cuerdas cuyas resistencias al rompimiento tienen media de 300 lb y desviación estándar de 24 lb. Se cree que gracias a un proceso recientemente desarrollado, la media de la resistencia al rompimiento puede incrementarse. (*a*) Diseñe una regla de decisión para rechazar el proceso antiguo con un nivel de significancia de 0.01 si se decide examinar 64 cuerdas. (*b*) Bajo la regla de decisión adoptada en (*a*), ¿cuál es la probabilidad de aceptar el proceso antiguo cuando de hecho el proceso nuevo ha incrementado la resistencia promedio al rompimiento en 310 lb? Suponga que la desviación estándar aún es 24 lb.

(*a*) Si μ es la resistencia promedio al rompimiento, queremos decidir entre las hipótesis

H_0: $\mu = 300$ lb, y el proceso nuevo es equivalente al antiguo

H_1: $\mu > 300$ lb, y el proceso nuevo es mejor que el antiguo

Para una prueba de una cola a un nivel de significancia de 0.01, tenemos la siguiente regla de decisión (*véase* figura 7-8):

(*1*) Rechazar H_0 si el valor z de la media de la resistencia al rompimiento es mayor que 2.33.

(*2*) De otra manera aceptar H_0.

Dado que

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - 300}{24/\sqrt{64}},\ \bar{X} = 300 + 3Z.$$ Entonces si $Z > 2.33$, $\bar{X} > 300 + 3(2.33) = 307.0$ lb.

Por consiguiente, las reglas de decisión anteriores se convierten en:

(*1*) Rechazar H_0 si la media de la resistencia al rompimiento de 64 cuerdas excede 307.0 lb.

(2) De otra manera aceptar H_0.

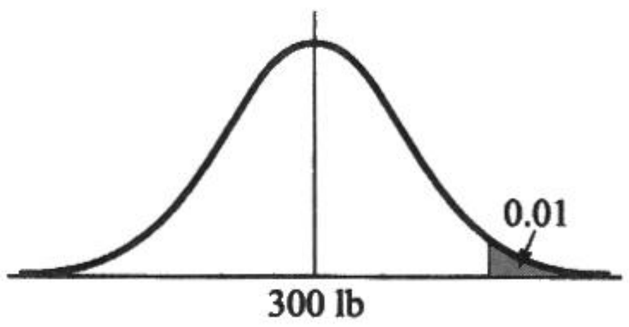


Fig. 7-8

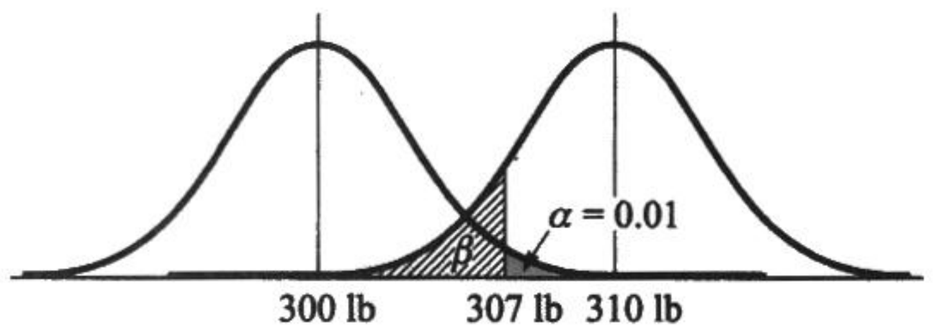


Fig. 7-9

(*b*) Considere las dos hipótesis (H_0: $\mu = 300$ lb) y (H_1: $\mu = 310$ lb). Las distribuciones de la resistencia al rompimiento correspondientes a estas dos hipótesis están representadas, respectivamente, por la parte izquierda y derecha de las distribuciones normales de la figura 7-9.

La probabilidad de aceptar el proceso antiguo cuando en realidad la media de la resistencia al rompimiento es 310 lb, está representada por la región de área β en la figura 7.9. Para encontrar esto tenga en cuenta que 307.0 lb en unidades estándar es $(307.0 - 310)/3 = -1.00$; dado que

β = área bajo la curva normal del lado derecho a la izquierda de $z = -1.00 = 0.1587$

Ésta es la probabilidad de aceptar (H_0: $\mu = 300$ lb) cuando en realidad (H_1: $\mu = 310$ lb) es verdadera; es decir, es la probabilidad de cometer un error de tipo II.

7.27. Construya (*a*) una curva OC, (*b*) una curva de potencia para el problema 7.26 suponiendo que la desviación estándar de la resistencia al rompimiento permanece en 24 lb.

Haciendo un razonamiento similar al del problema 7.26(b), podemos encontrar β para los casos en donde el proceso nuevo produce resistencias promedio al rompimiento μ iguales a 305 lb, 315 lb, etc. Por ejemplo, si $\mu = 305$ lb, entonces 307.0 lb en unidades estándar es $(307.0 - 305)/3 = 0.67$, y así

β = área bajo la curva normal del lado derecho a la izquierda de $z = 0.67 = 0.7486$

De esta manera se obtiene la tabla 7-4.

Tabla 7-4

μ	290	295	300	305	310	315	320
β	1.0000	1.0000	0.9900	0.7486	0.1587	0.0038	0.0000

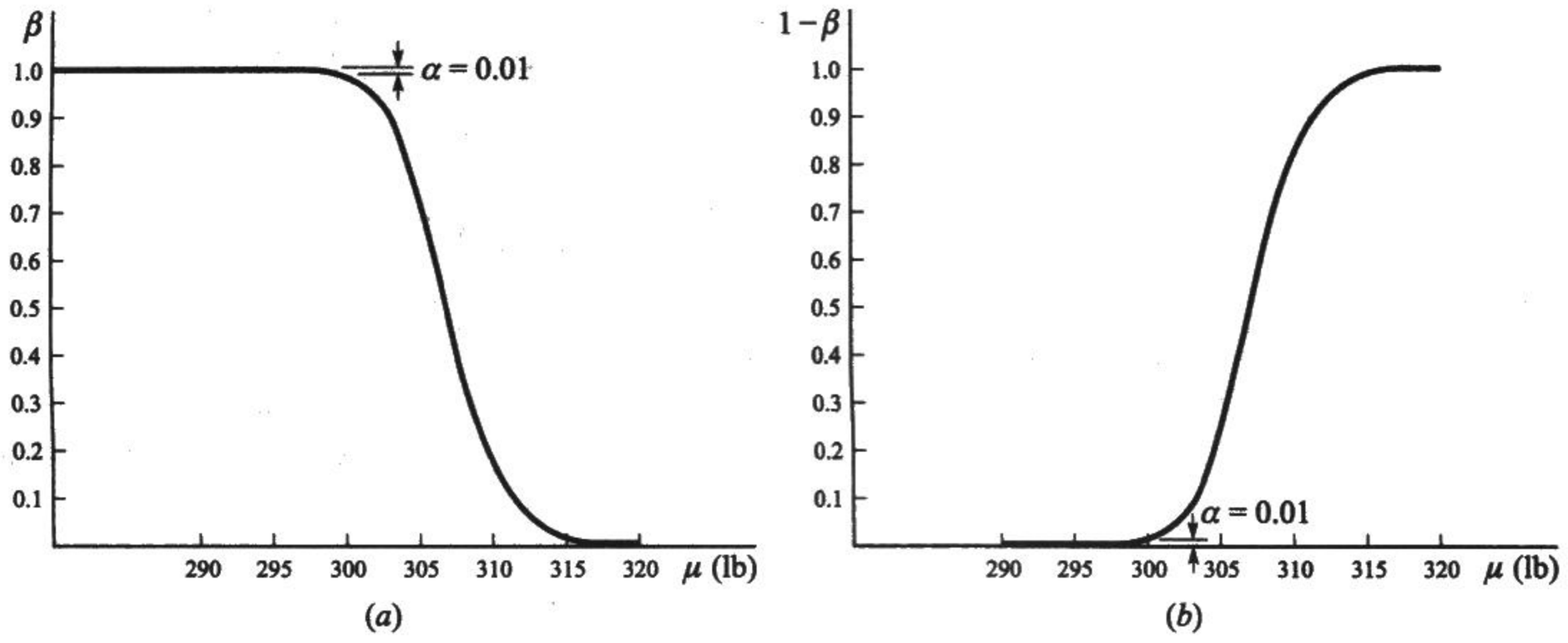


Fig. 7-10

(a) En la figura 7-10(a) se muestra la curva OC. De esta curva podemos ver que la probabilidad de mantener el proceso antiguo si la nueva resistencia al rompimiento es menor que 300 lb es prácticamente 1 (excepto por el nivel de significancia de 0.01 cuando el proceso nuevo tiene una media de 300 lb). Luego ésta cae abruptamente a cero, de manera que prácticamente no hay probabilidad de mantener el proceso antiguo cuando la resistencia media al rompimiento es mayor que 315 lb.

(b) La curva de potencia, que se muestra en la figura 7-10(b), es susceptible de la misma interpretación que la curva OC. De hecho, las dos curvas son esencialmente equivalentes.

7.28. Queremos imponer las siguientes restricciones en la prueba de la hipótesis de que una moneda es balanceada (es decir, $p = 0.5$) a partir de cierto número de lanzamientos de dicha moneda: (A) la probabilidad de rechazar la hipótesis cuando en realidad es correcta debe ser como máximo 0.05; (B) la probabilidad de aceptar la hipótesis cuando en realidad p difiere de 0.5 por 0.1 o más (es decir, $p \geqq 0.6$ o $p \leq 0.4$) debe ser como máximo 0.05. Determine el tamaño muestral mínimo necesario y plantee la regla de decisión resultante.

Aquí colocamos límites a los riesgos de tipo I y tipo II. Por ejemplo, las restricciones impuestas (A) requieren que la probabilidad de un error de tipo I sea $\alpha = 0.05$ como máximo, mientras que la restricción (B) requiere que la probabilidad de un error de tipo II sea $\beta = 0.05$ como máximo. La situación se ilustra gráficamente en la figura 7-11.

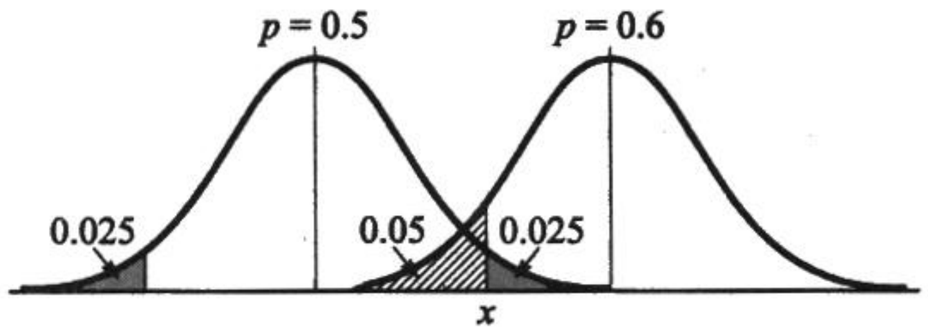


Fig. 7-11

Sea n el tamaño muestral requerido y x el número de caras en n lanzamientos por encima de los cuales rechazamos la hipótesis $p = 0.5$. De la figura 7-11,

(*1*) Área bajo la curva normal para $p = 0.5$ a la derecha de $\dfrac{x - np}{\sqrt{npq}} = \dfrac{x - 0.5n}{0.5\sqrt{n}}$ es 0.025.

(*2*) Área bajo la curva normal para $p = 0.6$ a la izquierda de $\dfrac{x - np}{\sqrt{npq}} = \dfrac{x - 0.6n}{0.49\sqrt{n}}$ es 0.05.

En realidad, equiparamos las áreas entre

$$\frac{(n - x) - 0.6n}{0.49\sqrt{n}} \qquad \text{y} \qquad \frac{x - 0.6n}{0.49\sqrt{n}}$$

a 0.05; sin embargo, (*2*) es una aproximación cercana. Observe que "en el peor de los casos" al hacer la probabilidad de aceptación 0.05, $p = 0.6$, automáticamente la convertimos en 0.05 o menor cuando p tiene cualquier otro valor por fuera del rango 0.4 a 0.6. Así, el promedio ponderado de todas estas probabilidades, las cuales representan la probabilidad de error de tipo II, también será 0.05 o menor.

De (*1*), $\dfrac{x - 0.5n}{0.5\sqrt{n}} = 1.96$ o (*3*) $x = 0.5n + 0.980\sqrt{n}$.

De (*2*), $\dfrac{x - 0.6n}{0.49\sqrt{n}} = -1.645$ o (*4*) $x = 0.6n - 0.806\sqrt{n}$.

Entonces de (*3*) y (*4*), $n = 318.98$. Se deduce que el tamaño muestral debe ser al menos 319, es decir, debemos lanzar la moneda al menos 319 veces. Colocando $n = 319$ en (*3*) o (*4*), $x = 177$.

Para $p = 0.5$, $x - np = 177 - 159.5 = 17.5$. Por consiguiente, adoptamos la siguiente regla de decisión:

(*a*) Aceptar la hipótesis $p = 0.5$ si el número de caras en 319 lanzamientos está en el rango 159.5 ± 17.5, es decir, entre 142 y 177.

(*b*) Rechazar la hipótesis de otra manera.

GRÁFICOS DE CONTROL DE CALIDAD

7.29. Una máquina está diseñada para producir balineras cuyo diámetro medio es 0.574 centímetros y desviación estándar de 0.008 centímetros. Para determinar si la máquina está funcionando adecuadamente, se tomó una muestra de 6 balineras cada 2 horas y a partir de ella se calculó la media del diámetro. (*a*) Diseñe la regla de decisión por medio de la cual se puede estar casi seguro de que la calidad del producto cumple con los estándares requeridos. (*b*) Muestre cómo se representa gráficamente la regla de decisión en (*a*).

(*a*) Con 99.73% de confianza podemos decir que la media muestral $\bar{X}$ debe estar en el rango $(\mu_{\bar{X}} - 3\sigma_{\bar{X}})$ a $(\mu_{\bar{X}} + 3\sigma_{\bar{X}})$ o $(\mu - 3\sigma/\sqrt{n})$ a $(\mu + 3\sigma/\sqrt{n})$. Dado que $\mu = 0.574$, $\sigma = 0.008$ y $n = 6$, se deduce que con 99.73% de confianza la media muestral deberá estar entre $(0.574 - 0.024/\sqrt{6})$ y $(0.574 + 0.024/\sqrt{6})$ o entre 0.564 y 0.584 centímetros.

Así, nuestra regla de decisión es como sigue:

(*1*) Si una media muestral cae dentro del rango de 0.564 a 0.584 centímetros, suponemos que la máquina está trabajando correctamente.

(2) De otra manera concluimos que la máquina no está trabajando adecuadamente y buscamos determinar la razón.

(*b*) Se puede mantener un registro de la media muestral por medio de un gráfico tal como el de la figura 7-12, llamado *gráfico de control de calidad*. Cada vez que se calcula una media, ésta se representa por un punto. En la medida en que los puntos se encuentren entre el límite inferior 0.564 y el superior 0.584, el proceso está bajo control. Cuando un punto se sale de estos límites de control (tal como el de la tercera muestra tomada el jueves), hay una posibilidad de que algo esté mal y se recomienda una investigación.

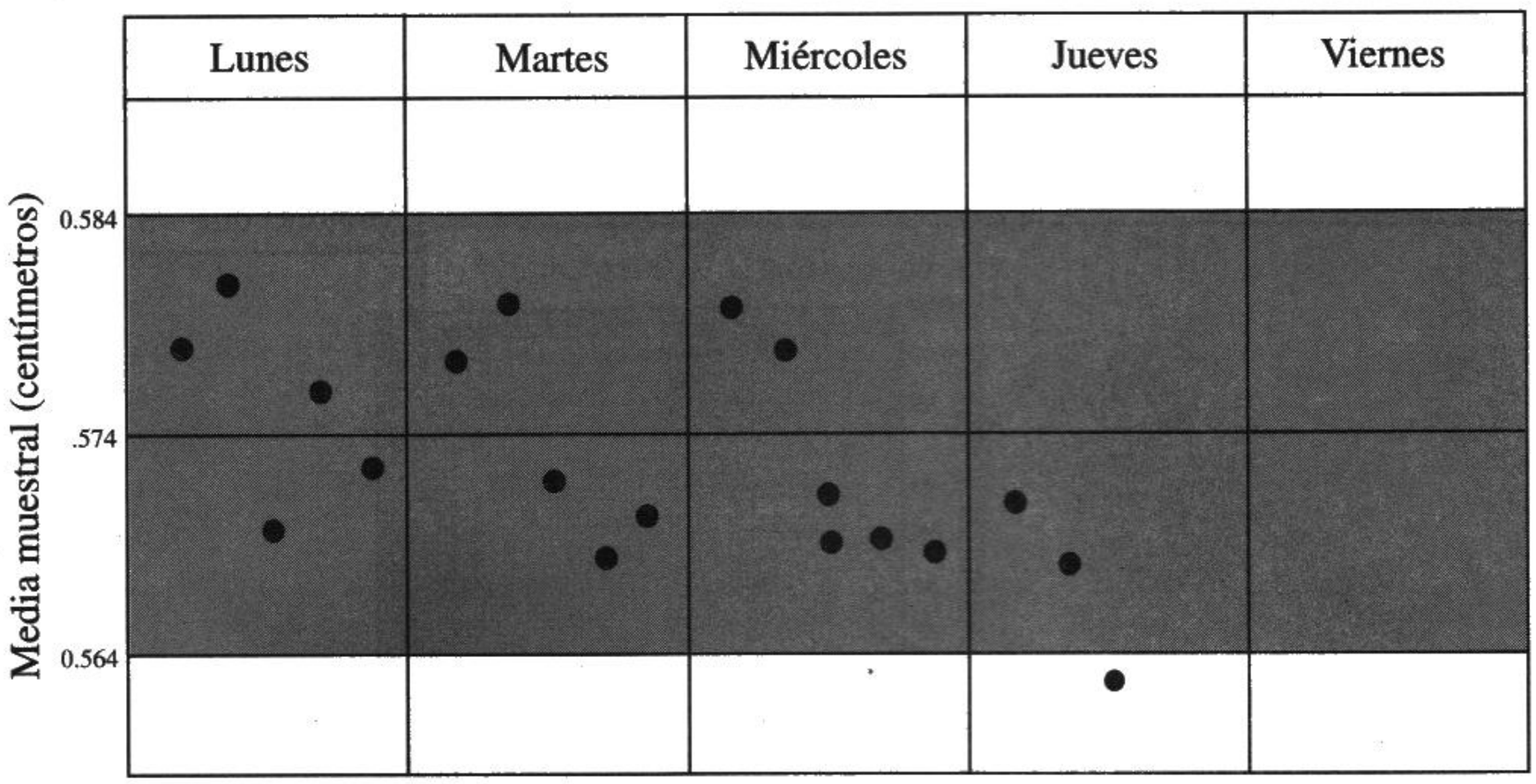


Fig. 7-12

Los límites de control especificados anteriormente se llaman límites de confianza del 99.73% o brevemente límites 3σ. Sin embargo, también se pueden determinar otros límites de confianza, tales como del 99% o 95%. En cada caso, la escogencia depende de las circunstancias particulares.

AJUSTE DE DATOS POR DISTRIBUCIONES TEÓRICAS

7.30. Ajuste una distribución binomial a los datos del problema 5.30, página 189.

Tenemos $P(x \text{ caras en un lanzamiento de 5 monedas}) = f(x) = {}_5C_x p^x q^{5-x}$, donde p y q son las probabilidades respectivas de una cara o un sello en el lanzamiento sencillo de una moneda. La media o el número esperado de caras es $\mu = np = 5p$.

Para la distribución verdadera u observada, el número promedio de caras es

$$\frac{\sum fx}{\sum f} = \frac{(38)(0) + (144)(1) + (342)(2) + (287)(3) + (164)(4) + (25)(5)}{1000} = \frac{2470}{1000} = 2.47$$

Igualando las medias teórica y observada, $5p = 2.47$ o $p = 0.494$. Por consiguiente, la distribución binomial ajustada está dada por $f(x) = {}_5C_x = (0.494)^x(0.506)^{5-x}$.

Estas probabilidades, así como las frecuencias observadas y esperadas (teóricas) se incluyen en la tabla 7-5. La bondad de ajuste se investiga en el problema 7.43.

7.31. Utilice papel gráfico de probabilidad para determinar si la distribución de frecuencia de la tabla 5-2, página 173, puede aproximarse cercanamente a una distribución normal.

Primero, la distribución de frecuencia dada se convierte en una distribución de frecuencia relativa acumulativa, como se muestra en la tabla 7-6. Luego, las frecuencias relativas acumulativas expresadas como porcentaje se dibujan contra los límites superiores de clase en papel gráfico especial de probabilidad, como se muestra en

la figura 7-13. El grado de cercanía a una línea recta que tenga los puntos dibujados, determina la bondad de ajuste de la distribución dada a una distribución normal. Aquí se ve que los datos se ajustan cercanamente a una distribución normal. *Véase* problema 7.32.

Tabla 7-5

Número de caras, x	$P(x$ caras)	Frecuencias esperadas	Frecuencias observadas
0	0.0332	33.2 o 33	38
1	0.1619	161.9 o 162	144
2	0.3162	316.2 o 316	342
3	0.3087	308.7 o 309	287
4	0.1507	150.7 o 151	164
5	0.0294	29.4 o 29	25

Tabla 7-6

Estatura (pulgadas)	Frecuencia relativa acumulativa (%)
Menores que 61.5	5.0
Menores que 64.5	23.0
Menores que 67.5	65.0
Menores que 70.5	92.0
Menores que 73.5	100.0

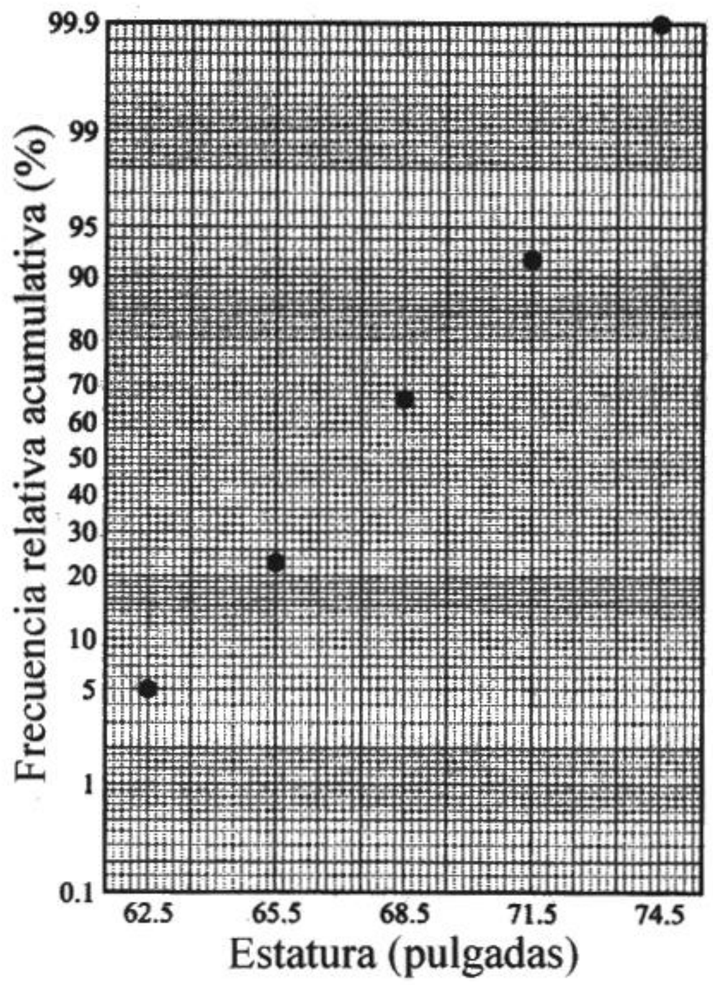


Fig. 7-13

7.32. Ajuste la curva normal a los datos de la tabla 5-2, página 173.

$$\bar{x} = 67.45 \text{ pulgadas}, \qquad s = 2.92 \text{ pulgadas}$$

El trabajo se puede organizar como en la tabla 7-7. Para calcular z para los límites de clase usamos $z = (x - \bar{x})/s$ donde la media $\bar{x}$ y la desviación estándar s se obtuvieron respectivamente en los problemas 5.35 y 5.40.

Tabla 7-7

Estatura (pulgadas)	Límites de clase x	z para límites de clase	Área bajo la curva normal de 0 a z	Área para cada clase	Frecuencia esperada	Frecuencia observada
60–62	59.5	−2.72	0.4967	0.0413	4.13 o 4	5
	62.5	−1.70	0.4554			
63–65				0.2086	20.68 o 21	18
	65.5	−0.67	0.2486			
66–68			} Suma	→ 0.3892	38.92 o 39	42
	68.5	0.36	0.1406			
69–71				0.2771	27.71 o 28	27
	71.5	1.39	0.4177			
72–74				0.0743	7.43 o 7	8
	74.5	2.41	0.4920			

En la cuarta columna, las áreas bajo la curva normal de 0 a z se obtuvieron usando la tabla del apéndice C. A partir de esto encontramos las áreas bajo la curva normal entre valores sucesivos de z, como en la quinta columna. Éstos se obtienen sustrayendo las áreas sucesivas de la cuarta columna cuando los z correspondientes tengan el mismo signo, y sumándolas cuando los z tengan signos opuestos (lo que ocurre sólo una vez en la tabla). La razón para esto se hace evidente en un diagrama.

Al multiplicar las entradas en la quinta columna (que representa las frecuencias relativas) por la frecuencia total n (en este caso $n = 100$) se obtienen las frecuencias esperadas o teóricas como en la sexta columna. Se ha visto que ellas concuerdan bien con las frecuencias verdaderas u observadas de la última columna.

La bondad de ajuste de la distribución se considera en el problema 7.44.

7.33. La tabla 7-8 muestra el número de días f de un periodo de 50 días durante el cual x accidentes automovilísticos ocurrieron en una ciudad. Ajuste los datos a una distribución de Poisson.

Tabla 7-8

Número de accidentes, x	Número de días, f
0	21
1	18
2	7
3	3
4	1
TOTAL	50

El número promedio de accidentes es

$$\lambda = \frac{\sum fx}{\sum f} = \frac{(21)(0) + (18)(1) + (7)(2) + (3)(3) + (1)(4)}{50}$$

$$= \frac{45}{50} = 0.90$$

Entonces, según la distribución de Poisson,

$$P(x \text{ accidentes}) = \frac{(0.90)^x e^{-0.90}}{x!}$$

En la tabla 7-9 aparecen las probabilidades para 0, 1, 2, 3 y 4 accidentes, como se obtuvieron de esta distribución de Poisson, así como el número teórico de días durante los cuales ocurrieron los z accidentes (se obtuvo multiplicando la probabilidad respectiva por 50). Para la conveniencia durante la comparación, se repitió la cuarta columna que da el número observado de días.

Tabla 7-9

Número de accidentes, x	$P(x$ accidentes)	Número esperado de días	Número observado de días
0	0.4066	20.33 ó 20	21
1	0.3659	18.30 ó 18	18
2	0.1647	8.24 ó 8	7
3	0.0494	2.47 ó 2	3
4	0.0111	0.56 ó 1	1

Observe que el ajuste de la distribución de Poisson a los datos es buena.

Para una distribución verdadera de Poisson, $\sigma^2 = \lambda$. El resultado de la variable de la distribución dada es 0.97. Ésta se compara favorablemente con el valor 0.90 para λ, lo cual puede tomarse como evidencia adicional para la idoneidad de la distribución de Poisson para aproximar los datos muestrales.

LA PRUEBA CHI CUADRADO

7.34. En 200 lanzamientos de una moneda se observaron 115 caras y 85 sellos. Pruebe la hipótesis de que la moneda es balanceada usando un nivel de significancia de (*a*) 0.05, (*b*) 0.01. (*c*) Encuentre el valor *P* de la prueba.

Las frecuencias observadas de caras y sellos son, respectivamente, $x_1 = 115$, $x_2 = 85$.

Las frecuencias esperadas de caras y sellos si la moneda es balanceada son $np_1 = 100$, $np_2 = 100$, respectivamente. Entonces

$$x^2 = \frac{(x_1 - np_1)^2}{np_1} + \frac{(x_2 - np_2)^2}{np_2} = \frac{(115 - 100)^2}{100} + \frac{(85 - 100)^2}{100} = 4.50$$

Dado que el número de categorías de clases (caras, sellos) es $k = 2$, $\nu = k - 1 = 2 - 1 = 1$.

(*a*) El valor crítico de $\chi^2_{0.95}$ para 1 grado de libertad es 3.84. Dado que $4.50 > 3.84$, rechazamos la hipótesis de que la moneda es balanceada al nivel de significancia de 0.05.

(*b*) El valor crítico de $\chi^2_{0.99}$ para 1 grado de libertad es 6.63. Ya que $4.50 < 6.63$, no podemos rechazar la hipótesis de que la moneda es balanceada al nivel de significancia de 0.01.

Concluimos que los resultados observados son *probablemente significativos* y que la moneda *probablemente no es balanceada*. Para una comparación de este método con otros métodos ya usados, *véase* el método 1 del problema 7.36.

(*c*) El valor *P* es $P(\chi^2 \geq 4.50)$. La tabla del apéndice E muestra $0.025 < P < 0.05$. Por computador, $P = 0.039$.

7.35. Desarrolle el problema 7.34 usando la corrección de Yates.

$$\chi^2 \text{ (corregida)} = \frac{(|x - 1 - np_1| - 0.5)^2}{np_1} + \frac{(|x_2 - np_2| - 0.5)^2}{np_2}$$

$$= \frac{(|115 - 100| - 0.5)^2}{100} + \frac{(|85 - 100| - 0.5)^2}{100} = \frac{(14.5)^2}{100} + \frac{(14.5)^2}{100} = 4.205$$

El valor *P* corregido es 0.04.

Dado que $4.205 > 3.84$ y $4.205 < 6.63$, las conclusiones a las que llegamos en el problema 7.34 son válidas. Para una comparación con métodos anteriores, *véase* el método 2 del problema 7.36.

7.36. Desarrolle el problema 7.34 usando la aproximación normal a la distribución binomial.

Bajo la hipótesis de que la moneda es balanceada, la media y la desviación estándar del número de caras en 200 lanzamientos de una moneda son $\mu = np = (200)(0.5) = 100$ y $\sigma = \sqrt{npq} = \sqrt{(200)(0.5)(0.5)} = 7.07$.

Método 1.

$$115 \text{ caras en unidades estándar} = (115 - 100)/7.07 = 2.12$$

Usando un nivel de significancia de 0.05 y una prueba de dos colas, rechazamos la hipótesis de que la moneda es balanceada si el valor z estuviera fuera del intervalo -1.96 y 1.96. Con un nivel de 0.01 el intervalo correspondiente sería -2.58 a 2.58. Se deduce como en el problema 7.34 que podemos rechazar la hipótesis con un nivel de 0.05, pero no podemos rechazarla con un nivel de 0.01. El valor *P* de la prueba es 0.034.

Observe que el cuadrado del valor estándar anterior, $(2.12)^2 = 4.50$, es el mismo valor de χ^2 obtenido en el problema 7.34. Éste es siempre el caso de una prueba chi cuadrado que involucra dos categorías (*véase* problema 7.60).

Método 2.

Usando la corrección para continuidad, 115 o más caras equivale a 114.5 o más caras. Entonces, 114.5 en unidades estándar $= (114.5 - 100)/7.07 = 2.05$. Esto nos lleva a la misma conclusión que en el primer método. El valor P corregido es 0.04.

Observe que el cuadrado de este valor estándar es $(2.05)^2 = 4.20$, el cual concuerda con el valor de χ^2 corregido para continuidad usando la corrección de Yates del problema 7.35. Éste siempre es el caso para una distribución chi cuadrado que involucra dos categorías en las que se aplica la corrección de Yates, de nuevo como consecuencia del problema 7.60.

7.37. La tabla 7-10 muestra las frecuencias esperada y observada al lanzar un dado 120 veces. (*a*) Pruebe la hipótesis de que el dado es balanceado usando un nivel de significancia de 0.05. (*b*) Encuentre el valor P de la prueba.

(*a*)

Tabla 7-10

Cara	1	2	3	4	5	6
Frecuencia observada	25	17	15	23	24	16
Frecuencia esperada	20	20	20	20	20	20

$$\chi^2 = \frac{(x_1 - np_1)^2}{np_1} + \frac{(x_2 - np_2)^2}{np_2} + \frac{(x_3 - np_3)^2}{np_3} + \frac{(x_4 - np_4)^2}{np_4} + \frac{(x_5 - np_5)^2}{np_5} + \frac{(x_6 - np_6)^2}{np_6}$$

$$= \frac{(25 - 20)^2}{20} + \frac{(17 - 20)^2}{20} + \frac{(15 - 20)^2}{20} + \frac{(23 - 20)^2}{20} + \frac{(24 - 20)^2}{20} + \frac{(16 - 20)^2}{20} = 5.00$$

Dado que el número de categorías o clases (caras 1, 2, 3, 4, 5, 6) es $k = 6$, $\nu = k - 1 = 5$.

El valor crítico de $\chi^2_{0.95}$ para 5 grados de libertad es 11.1. Como $5.00 < 11.1$, no rechazamos la hipótesis de que el dado está balanceado.

Para 5 grados de libertad $\chi^2_{0.05} = 1.15$; así que $\chi^2 = 5.00 > 1.15$. Se deduce que no concuerdan de manera excepcional, así que debemos ver el resultado con suspicacia.

(*b*) El valor P de la prueba es $P(\chi^2 > 5.00)$. La tabla del apéndice E muestra $0.25 < P < 0.5$. Por computador, $P = 0.42$.

7.38. Una tabla de números aleatorios de 250 dígitos tiene la distribución de los dígitos 0, 1, 2, . . ., 9 que muestra la tabla 7-11. (*a*) ¿Difiere significativamente la distribución observada de la distribución esperada? (*b*) ¿Cuál es el valor P de la observación?

(*a*)

Tabla 1-11

Dígito	0	1	2	3	4	5	6	7	8	9
Frecuencia observada	17	31	29	18	14	20	35	30	20	36
Frecuencia esperada	25	25	25	25	25	25	25	25	25	25

$$\chi^2 = \frac{(17-25)^2}{25} + \frac{(31-25)^2}{25} + \frac{(29-25)^2}{25} + \frac{(18-25)^2}{25} + \cdots + \frac{(36-25)^2}{25} = 23.3$$

El valor crítico $\chi^2_{0.99}$ para $\nu = k - 1 = 9$ grados de libertad es 21.7 y $23.3 > 21.7$. Así concluimos que la distribución observada difiere significativamente de la distribución esperada al nivel de significancia de 0.01. Sin embargo, conviene ser suspicaces.

(*b*) El valor P es $P(\chi^2 \geq 23.3)$. La tabla del apéndice E muestra que $0.005 < P < 0.01$. Por computador, $P = 0.0056$.

7.39. En sus experimentos con guisantes Mendel observó 315 redondos y amarillos, 108 redondos y verdes, 101 arrugados y amarillos y 32 arrugados y verdes. Según su teoría de la herencia, los números debían tener la proporción 9 : 3 : 3 :1. ¿Hay alguna evidencia para dudar de su teoría con un nivel de significancia de (*a*) 0.01, (*b*) 0.05? (*c*) ¿Cuál es el valor P de la observación?

El número total de guisantes es 315 + 108 + 101 + 32 = 556. Dado que el número esperado está en la proporción 9 : 3 : 3 :1 (y 9 + 3 + 3 +1 = 16), podemos esperar

$\frac{9}{16}(556) = 312.75$ redondos y amarillos $\quad \frac{3}{16}(556) = 104.25$ arrugados y amarillos

$\frac{3}{16}(556) = 104.25$ redondos y verdes $\quad \frac{1}{16}(556) = 34.75$ arrugados y verdes

Entonces

$$\chi^2 = \frac{(315-312.75)^2}{312.75} + \frac{(108-104.25)^2}{104.25} + \frac{(101-104.25)^2}{104.25} + \frac{(32-34.75)^2}{34.75} = 0.470$$

Dado que hay cuatro categorías, $k = 4$ y el número de grados de libertad es $\nu = 4 - 1 = 3$.

(*a*) Para $\nu = 3$, $\chi^2_{0.99} = 11.3$ de manera que no podemos rechazar la teoría al nivel de 0.01.

(*b*) Para $\nu = 3$, $\chi^2_{0.95} = 7.81$ de manera que no podemos rechazar la teoría al nivel de 0.05.

Concluimos que la teoría y el experimento concuerdan.

Observe que para 3 grados de libertad, $\chi^2_{0.05} = 0.352$ y $\chi^2 = 0.470 > 0.352$. Por consiguiente, a pesar de que la concordancia es buena, los resultados obtenidos están sujetos a una cantidad razonable de error muestral.

(*c*) El valor P es $P(\chi^2 = 0.470)$. La tabla del apéndice E muestra que $0.9 < P < 0.95$. Por computador, $P = 0.93$.

7.40. En una urna hay un número considerablemente grande de canicas de cuatro colores diferentes: rojo, anaranjado, amarillo y verde. Se toma al azar de la urna una muestra de 12 canicas, que resultaron ser 2 canicas rojas, 5 anaranjadas, 4 amarillas y 1 verde. Pruebe la hipótesis de que la urna contiene cantidades iguales de canicas de diferentes colores y encuentre el valor P de los resultados muestrales.

Bajo la hipótesis de que la urna contiene proporciones iguales de canicas de diferentes colores, podríamos esperar 3 canicas de cada clase en una muestra de 12 canicas.

Dado que estos números esperados son menores que 5, la aproximación de chi cuadrado sería un error. Para evitar esto, combinamos categorías de manera que el número esperado en cada categoría sea, al menos, 5.

Si quisiéramos rechazar la hipótesis, deberíamos combinar categorías de tal manera que la evidencia en contra de la hipótesis surja fácilmente. En nuestro caso esto se logra considerando las categorías "roja o verde" y "anaranjada o amarilla", para las cuales la muestra reveló 3 y 9 canicas, respectivamente. Ya que el número esperado en cada categoría bajo la hipótesis de proporciones iguales es 6, tenemos

$$\chi^2 = \frac{(3-6)^2}{6} + \frac{(9-6)^2}{6} = 3$$

Para $\nu = 2 - 1 = 1$, $\chi^2_{0.95} = 3.84$. Por consiguiente, no podemos rechazar la hipótesis al nivel de significancia de 0.05 (aunque sí podemos al nivel de 0.10). Es concebible que los resultados observados surjan al azar incluso cuando los colores están en proporciones iguales. El valor P es $P(\chi^2 \geq 3) = 0.083$.

Otro método.

Usando la corrección de Yates encontramos

$$\chi^2 = \frac{(|3-6|-0.5)^2}{6} + \frac{(|9-6|-0.5)^2}{6} = \frac{(2.5)^2}{6} + \frac{(2.5)^2}{6} = 2.1$$

lo cual conduce a la misma conclusión anterior. Por supuesto, esto es de esperarse ya que la corrección de Yates siempre *reduce* el valor de χ^2. Aquí el valor P es $P(\chi^2 > 2.1) = 0.15$.

Se debe tener en cuenta que si se usa la aproximación χ^2 a pesar del hecho de que las frecuencias son muy pequeñas, podremos obtener

$$\chi^2 = \frac{(2-3)^2}{3} + \frac{(5-3)^2}{3} + \frac{(4-3)^2}{3} + \frac{(1-3)^2}{3} = 3.33$$

con un valor P de 0.34.

Así que para $\nu = 4 - 1 = 3$, $\chi^2_{0.95} = 7.81$, llegamos a la misma conclusión anterior. Infortunadamente la aproximación χ^2 para frecuencias pequeñas es pobre; por tanto, cuando no es recomendable combinar frecuencias debemos recurrir a métodos de probabilidad exacta que involucre la distribución multinomial.

7.41. En 360 lanzamientos de un par de dados se observaron 74 "sietes" y 24 "onces". Usando un nivel de significancia de 0.05, pruebe la hipótesis de que los dados son balanceados y encuentre el valor P de los resultados observados.

Un par de dados puede caer de 36 maneras. Un siete puede ocurrir de 6 maneras y un once de 2 maneras.

Entonces $P(\text{siete}) = \frac{6}{36} = \frac{1}{6}$ y $P(\text{once}) = \frac{2}{36} = \frac{1}{18}$. Por consiguiente, en 360 lanzamientos podemos esperar $\left(\frac{1}{6}\right)(360) = 60$ sietes y $\left(\frac{1}{18}\right)(36) = 20$ onces, de manera que

$$\chi^2 = \frac{(74-60)^2}{60} + \frac{(24-20)^2}{20} = 4.07$$

con un valor P de 0.044.

Para $\nu = 2 - 1 = 1$, $\chi^2_{0.95} = 3.84$. Entonces, ya que $4.07 > 3.84$, estamos inclinados a rechazar la hipótesis de que los dados son balanceados. Sin embargo, usando la corrección de Yates encontramos

$$\chi^2 \text{ (corregida)} = \frac{(|74-60|-0.5)^2}{60} + \frac{(|24-20|-0.5)^2}{20} = \frac{(13.5)^2}{60} + \frac{(3.5)^2}{20} = 3.65$$

con un valor P de 0.056.

Por lo tanto, con base en la χ^2 corregida podemos no rechazar la hipótesis al nivel de 0.05.

En general, para muestras grandes, tal como la que tenemos aquí, usando la corrección de Yates los resultados demuestran ser más confiables que los resultados sin corrección. Sin embargo, ya que el valor corregido de χ^2 está muy cerca del valor crítico, estamos indecisos en cuanto a tomar decisiones en un sentido o en otro. En tales casos tal vez lo mejor sea aumentar el tamaño de la muestra tomando más observaciones, si estamos interesados especialmente en el nivel 0.05. De otra manera, podemos rechazar la hipótesis a algún otro nivel (tal como 0.10).

7.42. Una investigación de 320 familias con 5 hijos cada una reveló la distribución de niños y niñas como aparece en la tabla 7-12. (*a*) ¿Es consistente el resultado con la hipótesis de que los nacimientos de niños y niñas son igualmente probables? (*b*) ¿cuál es el valor *P* de los resultados muestrales?

(*a*)

Tabla 7-12

Número de niños y niñas	5 niños 0 niñas	4 niños 1 niña	3 niños 2 niñas	2 niños 3 niñas	1 niño 4 niñas	0 niños 5 niñas	TOTAL
Número de familias	18	56	110	88	40	8	320

Sea p = probabilidad de nacimiento de un niño y $q = 1 - p$ = probabilidad de nacimiento de una niña. Entonces las probabilidades de (5 niños), (4 niños y 1 niña), . . ., (5 niñas) están dadas en términos de la expansión binomial

$$(p+q)^5 = p^5 + 5p^4q + 10p^3q^2 + 10p^2q^3 + 5pq^4 + q^5$$

Si $p = q = \frac{1}{2}$, tenemos

$$P(\text{5 niños y 0 niñas}) = \left(\frac{1}{2}\right)^5 = \frac{1}{32} \qquad P(\text{2 niños y 3 niñas}) = 10\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^3 = \frac{10}{32}$$

$$P(\text{4 niños y 1 niña}) = 5\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right) = \frac{5}{32} \qquad P(\text{1 niño y 4 niñas}) = 5\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^4 = \frac{5}{32}$$

$$P(\text{3 niños y 2 niñas}) = 10\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2 = \frac{10}{32} \qquad P(\text{0 niños y 5 niñas}) = \left(\frac{1}{2}\right)^5 = \frac{1}{32}$$

Entonces el número esperado de familias con 5, 4, 3, 2, 1 y 0 niños se obtiene, respectivamente, multiplicando las probabilidades anteriores por 320 y los resultados son 10, 50, 100, 100, 50, 10. De ahí

$$\chi^2 = \frac{(18-10)^2}{10} + \frac{(56-50)^2}{50} + \frac{(100-100)^2}{100} + \frac{(88-100)^2}{100} + \frac{(40-50)^2}{50} + \frac{(8-10)^2}{10} = 12.0$$

Dado que $\chi^2_{0.95} = 11.1$ y $\chi^2_{0.99} = 15.1$ para $\nu = 6 - 1 = 5$ grados de libertad, podemos rechazar la hipótesis al nivel de significancia de 0.05, pero no al de 0.01. Por lo tanto, concluimos que los resultados son probablemente significativos, y que los nacimientos de niño y niña no son igualmente probables.

(*b*) El valor *P* es $P(\chi^2 \geq 12.0) = 0.035$

BONDAD DE AJUSTE

7.43. Use la prueba chi cuadrado para determinar la bondad de ajuste de los datos del problema 7.30.

$$\chi^2 = \frac{(38-33.2)^2}{33.2} + \frac{(144-161.9)^2}{161.9} + \frac{(342-316.2)^2}{316.2} + \frac{(287-308.7)^2}{308.7} + \frac{(164-150.7)^2}{150.7} + \frac{(25-29.4)^2}{29.4}$$
$$= 7.45$$

Dado que el número de parámetros usados para calcular las frecuencias esperadas es $m = 1$ (concretamente, el parámetro p de la distribución binomial), $\nu = k - 1 - m = 6 - 1 - 1 = 4$.

Para $\nu = 4$, $\chi^2_{0.95} = 9.49$. El ajuste de los datos es bueno.

Para $\nu = 4$, $\chi^2_{0.05} = 0.711$. Por consiguiente, dado que $\chi^2 = 7.54 > 0.711$, el ajuste no es tan bueno como increíble.

El valor *P* es $P(\chi^2 \geq 7.45) = 0.11$.

7.44. Determine la bondad de ajuste de los datos del problema 7.32.

$$\chi^2 = \frac{(5-4.13)^2}{4.13} + \frac{(18-20.68)^2}{20.68} + \frac{(42-38.92)^2}{38.92} + \frac{(27-27.71)^2}{27.71} + \frac{(8-7.43)^2}{7.43} = 0.959$$

Ya que el número de parámetros utilizado para calcular las frecuencias esperadas es $m = 2$ (concretamente, la media μ y la desviación estándar σ de la distribución normal), $\nu = k - 1 - m = 5 - 1 - 2 = 2$.

Para $\nu = 2$, $\chi^2_{0.95} = 5.99$. Por consiguiente, concluimos que el ajuste de los datos es muy bueno.

Para $\nu = 2$, $\chi^2_{0.05} = 0.103$. Entonces, dado que $\chi^2 = 0.959 > 0.103$, el ajuste no es tan bueno.

El valor P es $P(\chi^2 \geq 0.959) = 0.62$

TABLAS DE CONTINGENCIA

7.45. Desarrolle el problema 7.13 usando la prueba chi cuadrado.

Las condiciones del problema se presentan en la tabla 7-13. Bajo la hipótesis nula H_0 de que el suero no tiene efecto, podemos esperar que 70 personas se recuperen en cada grupo y 30 en cada grupo no lo hagan, como se indica en la tabla 7-14. Observe que H_0 equivale a decir que la recuperación es independiente del uso del suero, es decir, que las clasificaciones son independientes.

Tabla 7-13 Frecuencias observadas

	Se recuperan	No se recuperan	TOTAL
Grupo *A* (usando suero)	75	25	100
Grupo *B* (no usando suero)	65	35	100
TOTAL	140	60	200

Tabla 7-14 Frecuencias esperadas bajo H_0

	Se recuperan	No se recuperan	TOTAL
Grupo *A* (usando suero)	70	30	100
Grupo *B* (no usando suero)	70	30	100
TOTAL	140	60	200

$$\chi^2 = \frac{(75-70)^2}{70} + \frac{(65-70)^2}{70} + \frac{(25-30)^2}{30} + \frac{(35-30)^2}{30} = 2.38$$

Para determinar el número de grados de libertad, considere la tabla 7-15, la cual es la misma que las tablas 7-13 y 7-14, excepto que sólo se muestran los totales. Es claro que tenemos la libertad de colocar solamente un número en cualquiera de las celdas vacías, ya que una vez que esto está hecho, los números en las celdas restantes se determinan excepcionalmente a partir de los totales indicados. Por consiguiente, hay 1 grado de libertad.

Tabla 7-15

	Se recuperan	No se recuperan	TOTAL
Grupo *A*			100
Grupo *B*			100
TOTAL	140	60	200

Dado que $\chi^2_{0.95} = 3.84$ para 1 grado de libertad y que $\chi^2 = 2.38 < 3.84$, concluimos que los resultados no son significativos al nivel de 0.05. Por lo tanto, estamos en incapacidad de rechazar H_0 a este nivel, y concluimos que el suero no es efectivo o posponemos la toma de decisión hasta que hagamos otras pruebas. El valor *P* de las frecuencias observadas es $P(\chi^2 \geq 2.38) = 0.12$.

Observe que $\chi^2 = 2.38$ es el cuadrado del valor $z = 1.54$ obtenido en el problema 7.13. En general, la prueba chi cuadrado que involucra proporciones muestrales en una tabla de contingencia 2×2 es equivalente a la prueba de significancia de diferencias en proporciones usando la aproximación normal, como en la página 233.

Observe también que aquí el valor *P* 5 0.12 es el doble del valor *P*, 0.0618 en el problema 7.13. En general, una prueba de una cola que utilice χ^2 es equivalente a una prueba de dos colas que use χ^2, ya que por ejemplo, $\chi^2 . \chi^2_{0.95}$ corresponde a $\chi . \chi_{0.95}$ o $\chi , 2\chi_{0.95}$. Dado que para tablas 2 3 2 χ^2 es el cuadrado del valor *z*, para este caso χ es lo mismo que *z*. Por lo tanto, el rechazo de la hipótesis al nivel de 0.05, usando χ^2, es equivalente al rechazo en una prueba de dos colas al nivel de 0.01 usando *z*.

7.46. Desarrolle el problema 7.45 usando la corrección de Yates.

$$\chi^2 \text{(corregida)} = \frac{(|75-70|-0.5)^2}{70} + \frac{(|65-70|-0.5)^2}{70} + \frac{(|25-30|-0.5)^2}{30} + \frac{(|35-30|-0.5)^2}{30} = 1.93$$

con un valor *P* de 0.16.

Por consiguiente, las conclusiones obtenidas en el problema 7.45 son válidas. Hubiéramos podido darnos cuenta de esto directamente al observar que la corrección de Yates siempre disminuye el valor de χ^2 y aumenta el valor de *P*.

7.47. En la tabla 7-16 están indicados los números de estudiantes que ganaron y perdieron el curso según los tres profesores: Prof. *X*, Prof. *Y* y Prof. *Z*. Pruebe la hipótesis de que la proporción de estudiantes que pierde el curso con los tres profesores es la misma.

Tabla 7-16 Frecuencias observadas

	Prof. *X*	Prof. *Y*	Prof. *Z*	TOTAL
Ganaron	50	47	56	153
Perdieron	5	14	8	27
TOTAL	55	61	64	180

Bajo la hipótesis H_0 de que la proporción de estudiantes que pierden el curso con los tres profesores es la misma, habrían perdido $27/180 = 15\%$ de los estudiantes y habrían pasado el 85% de los estudiantes. Las frecuencias esperadas bajo H_0 se muestran en la tabla 7-17.

Entonces,

$$\chi^2 = \frac{(50-46.75)^2}{46.75} + \frac{(47-51.85)^2}{51.85} + \frac{(56-54.40)^2}{54.40} + \frac{(5-8.25)^2}{8.25} + \frac{(14-9.15)^2}{9.15} + \frac{(8-9.60)^2}{9.60} = 4.84$$

Tabla 7-17 Frecuencias esperadas bajo H_0

	Prof. X	Prof. Y	Prof. Z	TOTAL
Ganaron	85% de 55 = 46.75	85% de 61 = 51.85	85% de 64 = 54.40	153
Perdieron	15% de 55 = 8.25	15% de 61 = 9.15	15% de 64 = 9.60	27
TOTAL	55	61	64	180

Para determinar el número de grados de libertad, considere la tabla 7-18, que es la misma que las tablas 7-16 y 7-17, excepto porque sólo se muestran los totales. Es claro que tenemos la libertad de colocar sólo un número en una celda vacía de la primera columna y un número en una celda vacía de la segunda y tercera columna, después de lo cual todos los números en las celdas restantes se determinarán excepcionalmente a partir de los totales indicados. Por lo tanto, hay 2 grados de libertad en este caso.

Tabla 7-18

	Prof. X	Prof. Y	Prof. Z	TOTAL
Ganaron				153
Perdieron				27
TOTAL	55	61	64	180

Dado que $\chi^2_{0.95} = 5.99$, no podemos rechazar H_0 al nivel de 0.05. Sin embargo, observe que dado que $\chi^2_{0.90} = 4.61$, podemos rechazar H_0 al nivel de 0.10 si estamos dispuestos a asumir el riesgo de estar equivocados de 1 en 10. El valor P de las frecuencias observadas es $P(\chi^2 \geq 4.84) = 0.089$.

7.48. Demuestre que para una tabla de contingencia $h \times k$ $(h > 1, k > 1)$ el número de grados de libertad está dado por $(h-1)(k-1)$.

Hay $h + k - 1$ totales independientes de las hk entradas. Se deduce que el número de grados de libertad es

$$hk - (h+k-1) = (h-1)(k-1)$$

como se requiere. El resultado es válido si los parámetros de la población necesarios para obtener las frecuencias teóricas son conocidos; de otra manera, es necesario hacer ajustes, como se describió en (*b*), página 237.

7.49. La tabla 7-19 representa una tabla de contingencia general 2×2. Demuestre que

$$\chi^2 = \frac{n(a_1b_2 - a_2b_1)^2}{n_1 n_2 n_A n_B}$$

Tabla 7-19 Resultados observados

	I	II	TOTAL
A	a_1	a_2	n_A
B	b_1	b_2	n_B
TOTAL	n_1	n_2	n

Tabla 7-20 Resultados esperados

	I	II	TOTAL
A	$n_1 n_A/n$	$n_2 n_A/n$	n_A
B	$n_1 n_B/n$	$n_2 n_B/n$	n_B
TOTAL	n_1	n_2	n

Así como en el problema 7.45, los resultados esperados bajo la hipótesis nula se muestran en la tabla 7-20. Entonces,

$$\chi^2 = \frac{(a_1 - n_1 n_A/n)^2}{n_1 n_A/n} + \frac{(a_2 - n_2 n_A/n)^2}{n_2 n_A/n} + \frac{(b_1 - n_1 n_B/n)^2}{n_1 n_B/n} + \frac{(b_2 - n_2 n_B/n)^2}{n_2 n_B/n}$$

Pero
$$a_1 - \frac{n_1 n_A}{n} = a_1 - \frac{(a_1 + b_1)(a_1 + a_2)}{a_1 + b_1 + a_2 + b_2} = \frac{a_1 b_2 - a_2 b_1}{n}$$

De manera similar
$$a_2 - \frac{n_2 n_A}{n} = b_1 - \frac{n_1 n_B}{n} = b_2 - \frac{n_2 n_B}{n} = \frac{a_1 b_2 - a_2 b_1}{n}$$

Por consiguiente podemos escribir

$$\chi^2 = \frac{n}{n_1 n_A}\left(\frac{a_1 b_2 - a_2 b_1}{n}\right)^2 + \frac{n}{n_2 n_A}\left(\frac{a_1 b_2 - a_2 b_1}{n}\right)^2 + \frac{n}{n_1 n_B}\left(\frac{a_1 b_2 - a_2 b_1}{n}\right)^2 + \frac{n}{n_2 n_B}\left(\frac{a_1 b_2 - a_2 b_1}{n}\right)^2$$

lo cual simplifica a

(*1*)
$$\chi^2 = \frac{n(a_1 b_2 - a_2 b_1)^2}{n_1 n_2 n_A n_B} = \frac{n\Delta^2}{n_1 n_2 n_A n_B}$$

donde $\Delta = a_1 b_2 - a_2 b_1$, $n = a_1 + a_2 + b_1 + b_2$, $n_1 = a_1 + b_1$, $n_2 = a_2 + b_2$, $n_A = a_1 + a_2$, $n_B = b_1 + b_2$. Si se aplica la corrección de Yates, (*1*) se remplaza por

(*2*)
$$\chi^2(\text{corregida}) = \frac{n(|\Delta| - \frac{1}{2}n)^2}{n_1 n_2 n_A n_B}$$

7.50. Ilustre el resultado del problema 7.49 para los datos del problema 7.45.

En el problema 7.45, $a_1 = 75$, $a_2 = 25$; $b_1 = 65$, $b_2 = 35$, $n_1 = 140$, $n_2 = 60$, $n_A = 100$, $n_B = 100$ y $n = 200$; entonces, (*1*) del problema 7.49 da

$$\chi^2 = \frac{200[(75)(35) - (25)(65)]^2}{(140)(60)(100)(100)} = 2.38$$

Usando la corrección de Yates, el resultado es el mismo que el del problema 7.46:

$$\chi^2(\text{corregida}) = \frac{n(|a_1 b_2 - a_2 b_1| - \frac{1}{2}n)^2}{n_1 n_2 n_A n_B} = \frac{200[(75)(35) - (25)(65)| - 100]^2}{(140)(60)(100)(100)} = 1.93$$

7.51. Usando la aproximación normal (*véase* página 233) demuestre que una prueba chi cuadrado que involucra dos proporciones muestrales es equivalente a una prueba de significancia de diferencias en proporciones.

Sean P_1 y P_2 las dos proporciones muestrales y p los datos de la proporción de la población. En relación con el problema 7.49 tenemos

$$(1) \qquad P_1 = \frac{a_1}{n_1}, \qquad P_2 = \frac{a_2}{n_2}, \qquad 1 - P_1 = \frac{b_1}{n_1}, \qquad 1 - P_2 = \frac{b_2}{n_2}$$

$$(2) \qquad p = \frac{n_A}{n}, \qquad 1 - p = q = \frac{n_B}{n}$$

de manera que

$$(3) \qquad a_1 = n_1 P_1, \qquad a_2 = n_2 P_2, \qquad b_1 = n_1(1 - P_1), \qquad b_2 = n_2(1 - P_2)$$

$$(4) \qquad n_A = np, \qquad n_B = nq$$

Usando (*3*) y (*4*), tenemos del problema 7.49

$$\chi^2 = \frac{n(a_1 b_2 - a_2 b_1)^2}{n_1 n_2 n_A n_B} = \frac{n[n_1 P_1 n_2(1 - P_2) - n_2 P_2 n_1(1 - P_1)]^2}{n_1 n_2 npnq}$$

$$= \frac{n_1 n_2 (P_1 - P_2)^2}{npq} = \frac{(P_1 - P_2)^2}{pq(1/n_1 + 1/n_2)} \qquad \text{(ya que } n = n_1 + n_2\text{)}$$

lo cual es el cuadrado del estadístico Z dado en (10), página 233.

COEFICIENTE DE CONTINGENCIA

7.52. Encuentre el coeficiente de contingencia para los datos en la tabla de contingencia del problema 7.45.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{2.38}{2.38 + 200}} = \sqrt{0.01176} = 0.1084$$

7.53. Encuentre el valor máximo de C para todas las tablas 2×2 que puedan surgir en el problema 7.13.

El valor máximo de C ocurre cuando las dos clasificaciones son perfectamente dependientes o asociadas. En tales casos, todas las personas que tomen el suero se recuperarán y todas las que no lo tomen no se recuperarán. La tabla de contingencia se muestra como la tabla 7-21.

Tabla 7-21

	Se recuperan	No se recuperan	TOTAL
Grupo *A* (usando suero)	100	0	100
Grupo *B* (no usando suero)	0	100	100
TOTAL	100	100	200

Dado que las frecuencias esperadas de las celdas, suponiendo completa independencia, son iguales a 50,

$$\chi^2 = \frac{(100 - 50)^2}{50} + \frac{(0 - 50)^2}{50} + \frac{(0 - 50)^2}{50} + \frac{(100 - 50)^2}{50} = 200$$

Entonces el valor máximo de C es $\sqrt{\chi^2/(\chi^2 + n)} = \sqrt{200/(200 + 200)} = 0.7071$.

En general, para la perfecta dependencia en una tabla de contingencia donde el número de filas y de columnas es igual a k, las únicas frecuencias de las celdas que no tienen valor cero ocurren en la diagonal desde arriba a la izquierda hasta abajo a la derecha. Para tales casos, $C_{\text{máx}} = \sqrt{(k - 1)/k}$.

PROBLEMAS DIVERSOS

7.54. Un instructor hace una evaluación pequeña de 10 preguntas de falso verdadero. Para probar la hipótesis de que un estudiante está adivinando, se adoptó la siguiente regla de decisión: (*i*) Si 7 o más respuestas están correctas, el estudiante no está adivinando; (*ii*) si hay menos de 7 respuestas correctas, el estudiante está adivinando. Encuentre la probabilidad de rechazar la hipótesis cuando ésta es correcta.

Sea p = probabilidad de que la pregunta se responde correctamente.
La probabilidad de obtener x preguntas correctas de 10 es $10C_x p^x q^{10-x}$, donde $q = 1 - p$.
Entonces bajo la hipótesis $p = 0.5$ (es decir, el estudiante está adivinando),

$$P(\text{7 o más correctas}) = P(\text{7 correctas}) + P(\text{8 correctas}) + P(\text{9 correctas}) + P(\text{10 correctas})$$

$$= {}_{10}C_7\left(\frac{1}{2}\right)^7\left(\frac{1}{2}\right)^3 + {}_{10}C_8\left(\frac{1}{2}\right)^8\left(\frac{1}{2}\right)^2 + {}_{10}C_9\left(\frac{1}{2}\right)^9\left(\frac{1}{2}\right) + {}_{10}C_{10}\left(\frac{1}{2}\right)^{10} = \mathbf{0.1719}$$

Por consiguiente, la probabilidad de concluir que el estudiante no está adivinando cuando de hecho sí lo está haciendo es 0.1719. Observe que ésta es la probabilidad de error de tipo I.

7.55. En el problema 7.54, encuentre la probabilidad de aceptar la hipótesis $p = 0.5$ cuando en realidad $p = 0.7$.

Bajo la hipótesis $p = 0.7$,

$$P(\text{menos de 7 correctas}) = 1 - P(\text{7 o más correctas})$$

$$= 1 - [{}_{10}C_7(0.7)^7(0.3)^3 + {}_{10}C_8(0.7)^8(0.3)^2 + {}_{10}C_9(0.7)^9(0.3) + {}_{10}C_{10}(0.7)^{10}] = \mathbf{0.3504}$$

7.56. En el problema 7.54, encuentre la probabilidad de aceptar la hipótesis $p = 0.50$ cuando en realidad (*a*) $p = 0.6$, (*b*) $p = 0.8$, (*c*) $p = 0.9$, (*d*) $p = 0.4$, (*e*) $p = 0.3$, (*f*) $p = 0.2$, (*g*) $p = 0.1$.

(*a*) Si $p = 0.6$, la probabilidad requerida está dada por

$$1 - [P(\text{7 correctas}) + P(\text{8 correctas}) + P(\text{9 correctas}) + P(\text{10 correctas})]$$

$$= 1 - [{}_{10}C_7(0.6)^7(0.4)^3 + {}_{10}C_8(0.6)^8(0.4)^2 + {}_{10}C_9(0.6)^9(0.4) + {}_{10}C_{10}(0.6)^{10}] = \mathbf{0.618}$$

Los resultados para (*b*), (*c*), . . ., (*g*), se pueden encontrar de manera similar, y se indican en la tabla 7-22 junto con los valores correspondientes de $p = 0.7$ encontrados en el problema 7.55. Observe que la probabilidad se denota por β (probabilidad de tipo II). También incluimos la entrada para $p = 0.5$, dada por $\beta = 1 - 0.1719 = 0.828$ del problema 7.54.

Tabla 7-22

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
β	1.000	0.999	0.989	0.945	0.828	0.618	0.350	0.121	0.013

7.57. Use el problema 7.56 para construir una gráfica de b contra p, la curva de operación característica de la regla de decisión del problema 7.54.

La gráfica requerida se muestra en la figura 7-14. Observe que hay similitud con la curva *OC* del problema 7.27.

Si hubiéramos dibujado $(1 - \beta)$ contra p, se hubiese obtenido la *potencia de la curva* de la regla de decisión.

La gráfica indica que la regla de decisión dada tiene la potencia para rechazar $p = 0.5$ cuando en realidad $p \geqq 0.8$.

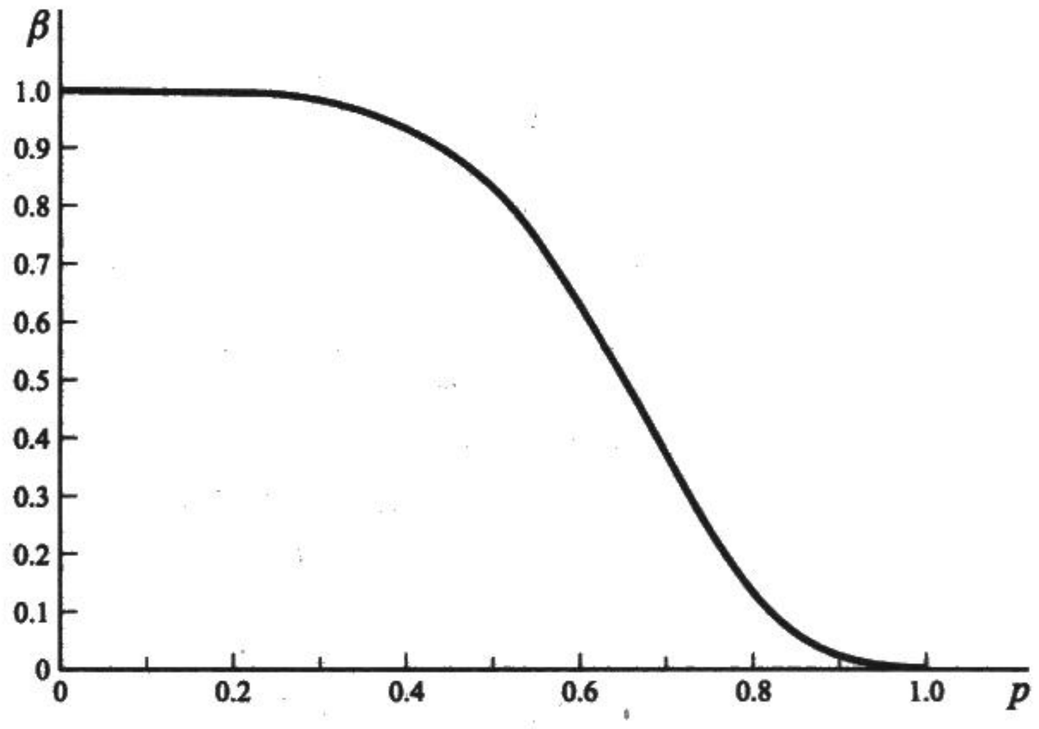


Fig. 7-14

7.58. Se lanza una moneda 6 veces y en todas aparece cara. ¿Podemos concluir al nivel de significancia de (*a*) 0.05, (*b*) 0.01 que la moneda no es balanceada? Considere pruebas de una y dos colas.

Sea p = probabilidad de caras en un lanzamiento único de la moneda.

Bajo la hipótesis (H_0: $p = 0.5$) (es decir, la moneda es balanceada),

$$f(x) = P(x \text{ caras en 6 lanzamientos}) = {}_6C_x\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{6-x} = \frac{{}_6C_x}{64}$$

Entonces la probabilidad de 0, 1, 2, 3, 4, 5 y 6 caras está dada respectivamente por $\frac{1}{64}, \frac{6}{64}, \frac{15}{64}, \frac{20}{64}, \frac{15}{64}, \frac{6}{64}$ y $\frac{1}{64}$.

Prueba de una cola.

Aquí queremos decidir entre las dos hipótesis (H_0: $p = 0.5$) y (H_1: p > 0.5). Ya que P(6 caras) $= \frac{1}{64} =$ 0.01562 y P(5 caras o 6 caras) $= \frac{6}{64} + \frac{1}{64} = 0.1094$, podemos rechazar H_0 al 0.05, pero no al nivel de 0.01 (o sea, los resultados observados son significativos al nivel de 0.05, pero no al nivel de 0.01).

Prueba de dos colas.

Aquí queremos decidir entre las hipótesis (H_0: $p = 0.5$) y (H_1: p $\neq 0.5$). Dado que P(0 ó 6 caras) $= \frac{1}{64} + \frac{1}{64} =$ 0.03125, podemos rechazar H_0 al nivel de 0.05, pero no al nivel de 0.01.

7.59. Desarrolle el problema 7.58 si en el lanzamiento de la moneda aparecen 5 caras.

Prueba de una cola.

Dado que P(5 ó 6 caras) $= \frac{6}{64} + \frac{1}{64} = \frac{7}{64} = 0.1094$, no podemos rechazar H_0 al nivel de 0.05 ni al nivel de 0.01.

Prueba de dos colas.

Ya que P(0 ó 1 ó 5 ó 6 caras) $= 2(\frac{7}{64}) = 0.2188$, no podemos rechazar H_0 al nivel de 0.05 ni al nivel de 0.01.

7.60. Demuestre que una prueba chi cuadrado que involucra dos categorías es equivalente a una prueba de significancia para proporciones (*véase* página 232).

Si P es la proporción muestral para la categoría I, p es la proporción poblacional, y n es la frecuencia total, podemos describir la situación por medio de la tabla 7.23. Luego, por definición,

$$\chi^2 = \frac{(nP - np)^2}{np} + \frac{[n(1-P) - n(1-p)]^2}{nq}$$

$$= \frac{n^2(P-p)^2}{np} + \frac{n^2(P-p)^2}{nq} = n(P-p)^2\left(\frac{1}{p} + \frac{1}{q}\right) = \frac{n(P-p)^2}{pq} = \frac{(P-p)^2}{pq/n}$$

que es el cuadrado del estadístico Z (*5*) de la página 232.

Tabla 7-23

	I	II	TOTAL
Frecuencia observada	nP	$n(1-P)$	n
Frecuencia esperada	np	$n(1-p)=nq$	n

7.61. Supongamos que $X_1, X_2, \ldots, X_k$ tienen distribución multinomial con frecuencias esperadas $np_1, np_2, \ldots, np_k$, respectivamente. Sean $Y_1, Y_2, \ldots, Y_k$ variables con distribución de Poisson mutuamente independientes, con parámetros $\lambda_1 = pn_1, \lambda_2 = np_2, \ldots, \lambda_k = np_k$, respectivamente. Demuestre que la distribución condicional de las Y dado que

$$Y_1 + Y_2 + \cdots + Y_k = n$$

es precisamente la distribución multinomial de las X.

Para la función de probabilidad conjunta de las Y, tenemos

$$(1) \qquad P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_k = y_k) = \left[\frac{(np_1)^{y_1}e^{-np_1}}{y_1!}\right]\left[\frac{(np_2)^{y_2}e^{-np_2}}{y_2!}\right]\cdots\left[\frac{(np_k)^{y_k}e^{-np_k}}{y_k!}\right]$$

$$= \frac{n^{y_1+y_2+\cdots+y_k}p_1^{y_1}p^{y_2}\cdots p_k^{y_k}}{y_1!y_2!\cdots y_k!}e^{-n}$$

donde usamos el hecho de que $p_1 + p_2 + \cdots + p_k = 1$. La distribución condicional que estamos buscando está dada por

$$(2) \quad P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_k = y_k \mid Y_1 + Y_2 + \cdots + Y_k = n)$$

$$= \frac{P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_k = y_k \text{ and } Y_1 + Y_2 + \cdots + Y_k = n)}{P(Y_1 + Y_2 + \cdots + Y_k = n)}$$

Ahora, a partir de (*1*) el numerador en (*2*) tiene el valor

$$\frac{n^n p_1^{y_1} p_2^{y_2}\cdots p_k^{y_k}}{y_1!y_2!\cdots y_k!}e^{-n}$$

Para el denominador, sabemos del problema 4.94, página 159, que $Y_1 + Y_2 + \cdots + Y_k$ es en sí misma una variable de Poisson con parámetro $np_1 + np_2 + \cdots + np_k = n$. Ya que el denominador tiene el valor

$$\frac{n^n e^{-n}}{n!}$$

Por consiguiente, (*2*) se convierte en

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_k = y_k \mid Y_1 + Y_2 + \cdots + Y_k = n) = \frac{n!}{y_1!y_2!\cdots y_k!}p_1^{y_1}p_2^{y_2}\cdots p_k^{y_k}$$

que es la distribución multinomial de la X [compare con (*16*), página 122].

7.62. Use el resultado del problema 7.61 para demostrar que χ^2, tal como se definió por (*21*), página 237, tiene aproximadamente la distribución chi cuadrado.

Tal como se establece, es difícil tratar con (*21*) porque las X con distribución binomial son independientes, según la restricción en (*22*). Sin embargo, el problema 7.61 muestra que podemos remplazar las $\times$ por las Y *independientes* que tienen distribución de Poisson, si está dado que $Y_1 + Y_2 + \cdots + Y_k = n$. Por lo tanto, podemos reescribir (*21*) como

$$(1) \qquad \chi^2 = \left(\frac{Y_1 - \lambda_1}{\sqrt{\lambda_1}}\right)^2 + \left(\frac{Y_2 - \lambda_2}{\sqrt{\lambda_2}}\right)^2 + \cdots + \left(\frac{Y_k - \lambda_k}{\sqrt{\lambda_k}}\right)^2$$

A medida que $n \to \infty$, todas las λ se aproximan a ∞, y el teorema del límite central para la distribución de Poisson [(*14*), página 120] da

$$(2) \qquad \chi^2 \approx Z_1^2 + Z_2^2 + \cdots + Z_k^2$$

donde las Z son variables *independientes* normales con media 0 y varianza 1, cuya distribución es condicional al evento

$$(3) \qquad \sqrt{\lambda_1}Z_1 + \sqrt{\lambda_2}Z_2 + \cdots + \sqrt{\lambda_k}Z_k = 0 \qquad \text{o} \qquad \sqrt{p_1}Z_1 + \sqrt{p_2}Z_2 + \cdots + \sqrt{p_k}Z_k = 0$$

o, dado que las variables aleatorias son continuas,

$$(4) \qquad |\sqrt{p_1}Z_1 + \sqrt{p_2}Z_2 + \cdots + \sqrt{p_k}Z_k| < \epsilon$$

Sea $F_\nu(x)$ la función de distribución acumulativa para una variable chi cuadrado con ν grados de libertad. Entonces lo que queremos demostrar es

$$(5) \qquad P(Z_1^2 + Z_2^2 + \cdots + Z_k^2 \leqq x \mid |\sqrt{p_1}Z_1 + \sqrt{p_2}Z_2 + \cdots + \sqrt{p_k}Z_k| < \epsilon)$$

$$= \frac{P(Z_1^2 + Z_2^2 + \cdots + Z_k^2 \leqq x \quad \text{y} \quad |\sqrt{p_1}Z_1 + \sqrt{p_2}Z_2 + \cdots + \sqrt{p_k}Z_k| < \epsilon)}{P(|\sqrt{p_1}Z_1 + \sqrt{p_2}Z_2 + \cdots + \sqrt{p_k}Z_k| < \epsilon)}$$

$$= F_\nu(x)$$

para un valor adecuado de ν.

Es fácil establecer (*5*) si usamos nuestra intuición geométrica. Primero que todo, el teorema 3-4 demuestra que la distribución incondicional de $Z_1^2 + Z_2^2 + \cdots + Z_k^2$ es chi cuadrado con k grados de libertad. Por lo tanto, ya que la función de densidad para cada Z_j es $(2\pi)^{-1/2}e^{-z^2/2}$,

$$(6) \qquad F_k(x) = (2\pi)^{-k/2} \int\cdots\int_{z_1^2+z_2^2+\cdots+z_k^2 \leqq x} e^{-(z_1^2+z_2^2+\cdots+z_k^2)/2}\, dz_1\, dz_2 \cdots dz_k$$

Además, para el numerador en (*5*) tenemos:

$$(7) \qquad \text{Numerador} = (2\pi)^{-k/2} \int\cdots\int_{\substack{z_1^2+z_2^2+\cdots+z_k^2 \leqq x,\\ |\sqrt{p_1}z_1+\sqrt{p_2}z_2+\cdots+\sqrt{p_k}z_k|<\epsilon}} e^{-(z_1^2+z_2^2+\cdots+z_k^2)}/2\, dz_1\, dz_2 \cdots dz_k$$

De la geometría analítica recordamos que en un espacio tridimensional $x_1^2 + x_2^2 + x_3^2 \leq a^2$ representa una esfera sólida con radio a centrado en el origen, mientras que $\alpha_1x_1 + \alpha_2x_2 + \alpha_3x_3 = 0$ es un plano que pasa a través del origen cuya normal es el vector $(\alpha_1, \alpha_2, \alpha_3)$. La figura 7-15 muestra la intersección de ambos cuerpos. Es obvio que cuando una función, que depende sólo de la distancia desde el origen, es decir,

$$f(r) \qquad \text{donde} \qquad r = \sqrt{x_1^2 + x_2^2 + x_3^2}$$

se integre sobre el área circular — o a través de una sección delgada ubicada en esa área — el valor de la integral es completamente independiente de la dirección de los cosenos α_1, α_2, α_3. En otras palabras, todos los planos que atraviesan el origen dan la misma integral.

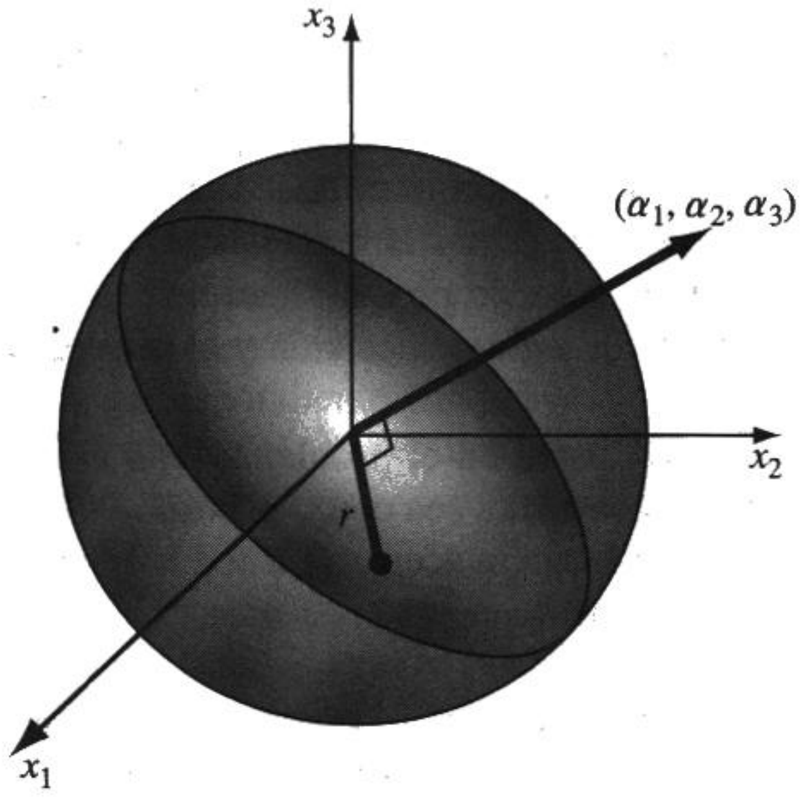


Fig. 7-15

Por analogía concluimos que en (7), donde $e^{-r^2/2}$ se integra sobre la intersección de un *hiperplano* a través del origen, se puede dar cualquier valor conveniente a las p. Escogemos

$$p_1 = p_2 = \cdots = p_{k-1} = 0, \qquad p_k = 1$$

y obtenemos

$$(8) \quad \text{Numerador} = (2\pi)^{-k/2} \int\cdots\int_{z_1^2+z_2^2+\cdots+z_{k-1}^2 \leq x} e^{-(z_1^2+z_2^2+\cdots+z_{k-1}^2)/2}\, dz_1\, dz_2 \cdots dz_{k-1}(2\epsilon)$$

$$= (2\pi)^{-1/2} F_{k-1}(x)(2\epsilon)$$

usando (6). El factor 2ϵ es el grosor de la sección.

Para evaluar el denominador en (5) observamos que la variable aleatoria

$$W = \sqrt{p_1}Z_1 + \sqrt{p_2}Z_2 + \cdots + \sqrt{p_k}Z_k$$

es normal (porque es una combinación lineal de las Z normales independientes), y que

$$E(W) = \sqrt{p_1}(0) + \sqrt{p_2}(0) + \cdots + \sqrt{p_k}(0) = 0$$
$$\text{Var}\,(W) = p_1(1) + p_2(1) + \cdots + p_k(1) = 1$$

Por consiguiente, la función de densidad para W es $\phi(w) = (2\pi)^{-1/2}e^{-w^2/2}$, y

$$(9) \qquad \text{Denominador} = P(|W| < \epsilon) = \phi(0)(2\epsilon) = (2\pi)^{-1/2}(2\epsilon)$$

Dividiendo (8) por (9), obtenemos el resultado deseado, donde $\nu = k - 1$.

La "prueba" anterior (la cual se puede hacer rigurosa) muestra, por cierto, que cada restricción lineal hecha a las Z, y de allí a las Y o a las X, reduce el número de grados de libertad en χ^2 en 1. Esto provee las bases para las reglas dadas en la página 237.

Problemas suplementarios

PRUEBAS PARA MEDIAS Y PROPORCIONES USANDO DISTRIBUCIONES NORMALES

7.63. Una urna contiene canicas rojas y azules. Para probar la hipótesis de que los colores se encuentran en proporciones iguales, decidimos tomar una muestra con remplazo de 64 canicas, anotando los colores que salen y adoptando la siguiente regla de decisión : (*1*) aceptar la hipótesis si se sacan entre 28 y 36 canicas rojas; (*2*) rechazar la hipótesis de otra manera. (*a*) Encuentre la probabilidad de rechazar la hipótesis cuando en realidad ésta es correcta. (*b*) Interprete gráficamente la regla de decisión y el resultado obtenido en (*a*).

7.64. (*a*) ¿Qué regla de decisión adoptaría usted en el problema 7.63 si necesita que la probabilidad de rechazar la hipótesis, cuando en realidad ésta es correcta, sea como máximo 0.01, o sea, usted quiere un nivel de significancia de 0.01? (*b*) ¿A qué nivel de significancia aceptaría usted la hipótesis? (*c*) ¿Cuál sería la regla de decisión si se toma un nivel de significancia de 0.05?

7.65. Supongamos que en el problema 7.63 usted quiere probar la hipótesis de que la proporción de canicas rojas es mayor que la de canicas azules. (*a*) ¿Cuál sería su hipótesis nula y cuál su hipótesis alterna? (*b*) ¿Usaría usted una prueba de una o dos colas? ¿Por qué? (*c*) ¿Qué regla de decisión adoptaría usted si el nivel de significancia fuera de 0.05? (*d*) ¿Cuál sería la regla de decisión si el nivel de significancia fuera 0.01?

7.66. Se lanza un par de dados 100 veces y se observa que aparecen "sietes" 23 veces. Pruebe la hipótesis de que los dados son balanceados, es decir, que no están cargados, usando (*a*) una prueba de dos colas y (*b*) una prueba de una cola, ambas con nivel de significancia de 0.05. Discuta las razones, si las hay, por las cuales prefirió una de las pruebas sobre la otra.

7.67. Desarrolle el problema 7.66 si el nivel de significancia es 0.01.

7.68. Un fabricante sostiene que al menos el 95% de los equipos que envió a una fábrica está acorde con las especificaciones. Una revisión de una muestra de 200 piezas reveló que 18 eran defectuosas. Pruebe la afirmación del fabricante al nivel de significancia de (*a*) 0.01, (*b*) 0.05.

7.69. La experiencia ha demostrado que la resistencia promedio al rompimiento de una marca particular de hilos es 9.72 onzas con desviación estándar de 1.4 oz. Recientemente una muestra de 36 piezas de hilo mostraron una resistencia promedio al rompimiento de 8.93 oz. ¿Se puede concluir que el hilo ha bajado de calidad con un nivel de significancia de (*a*) 0.05, (*b*) 0.01?

7.70. En un examen dado a estudiantes en un número grande de diferentes colegios se obtuvo una nota promedio de 74.5 con desviación estándar de 8.0. En cierto colegio, en donde 200 alumnos realizaron el examen, se obtuvo una nota promedio de 75.9. Discuta la significancia de este resultado a un nivel de 0.05 desde el punto de vista de (*a*) una prueba de una cola, (*b*) una prueba de dos colas, explicando cuidadosamente sus conclusiones con base en dichas pruebas.

7.71. Conteste el problema 7.70 si el nivel de significancia es 0.01.

PRUEBAS QUE INVOLUCRAN DIFERENCIAS DE MEDIAS Y PROPORCIONES

7.72. Una muestra de 100 bombillas eléctricas producidas por el fabricante *A* mostró tener un tiempo de vida media de 1190 horas y desviación estándar de 90 horas. Una muestra de 75 bombillas eléctricas producidas por el fabricante *B* mostró tener un tiempo de vida media de 1230 horas con desviación estándar de 120 horas. ¿Hay alguna diferencia entre los tiempos de vida media de las dos marcas de bombillas al nivel de significancia de (*a*) 0.05, (*b*) 0.01?

7.73. Usando un nivel de significancia de (*a*) 0.05, (*b*) 0.01 pruebe la hipótesis de que las bombillas del fabricante *B*, en el problema 7.72, tienen calidad superior que las del fabricante *A*. ¿Contradicen estos resultados los obtenidos en el problema 7.72?

7.74. En el examen de ortografía en un colegio, la nota promedio de 32 niños fue 72, con desviación estándar de 8, mientras que la nota promedio de las 36 niñas fue 75, con desviación estándar de 6. Con un nivel de significancia (*a*) 0.05, (*b*) 0.01 pruebe la hipótesis de que las niñas son mejores en ortografía que los niños.

7.75. Para probar el efecto de un nuevo fertilizante en la producción de trigo, una porción de terreno se dividió en 60 parcelas de áreas iguales, todas con idéntica calidad de suelo, exposición a la radiación solar, etc. El fertilizante nuevo se aplicó en 30 parcelas y el fertilizante viejo se aplicó a las parcelas restantes. El número promedio de toneladas cosechadas de trigo por parcela en donde se usó el fertilizante nuevo fue de 18.2, con desviación estándar de 0.63 toneladas. La media correspondiente y la desviación estándar de las parcelas a las que se aplicó el fertilizante viejo fueron 17.8 y 0.54 toneladas, respectivamente. Con un nivel de significancia de (*a*) 0.05, (*b*) 0.01, pruebe la hipótesis de que el fertilizante nuevo es mejor que el viejo.

7.76. Las muestras aleatorias de 200 tornillos fabricados por la máquina *A* y 100 fabricados por la máquina *B* mostraron 19 y 5 tornillos defectuosos, respectivamente. Pruebe la hipótesis de que (*a*) las dos máquinas están mostrando niveles diferentes de desempeño, (*b*) la máquina *B* se está desempeñando mejor que la *A*. Use un nivel de significancia de 0.05.

PRUEBAS QUE INVOLUCRAN LA DISTRIBUCIÓN *t* DE STUDENT

7.77. En el pasado, el tiempo de vida media de las bombillas eléctricas producidas por una compañía fue de 1120 horas con desviación estándar de 125 horas. Una muestra de 8 bombillas eléctricas tomadas de la producción reciente mostró una vida media de 1070 horas. Pruebe la hipótesis de que el tiempo de vida media de las bombillas es el mismo. Use un nivel de significancia de (*a*) 0.05, (*b*) 0.01.

7.78. En el problema 7.77 pruebe la hipótesis $\mu = 1120$ horas contra la hipótesis alterna $\mu < 1120$ horas, con un nivel de significancia de (*a*) 0.05, (*b*) 0.01.

7.79. Las especificaciones para la producción de cierta aleación requieren del 23.2% de cobre. Una muestra de 10 análisis del producto mostró un contenido medio de cobre del 23.5%, con desviación estándar de 0.24. ¿Podemos concluir con un nivel de significancia de (*a*) (0.01), (*b*) 0.05 que el producto cumple con las especificaciones requeridas?

7.80. En el problema 7.79 pruebe que la hipótesis de que la media del contenido de cobre es mayor que las especificaciones requeridas con un nivel de significancia de (*a*) 0.01, (*b*) 0.05.

7.81. Un experto en eficiencia sostiene que si se introduce maquinaria nueva en un proceso de producción, él puede disminuir sustancialmente el tiempo de producción. Debido a los altos costos de mantenimiento de las máquinas, la administración dice que a menos que el tiempo de producción se reduzca en un 8%, ellos no pueden pagar la introducción de las máquinas nuevas. Seis experimentos mostraron que el tiempo de producción se redujo en un 8.4% con desviación estándar de 0.32%. Pruebe la hipótesis de que se deben introducir las máquinas con un nivel de significancia de (*a*) 0.01, (*b*) 0.05.

7.82. Se probaron dos tipos de soluciones químicas, *A* y *B*, para probar su pH (el grado de acidez de la solución). El análisis de 6 muestras de *A* mostró un pH promedio de 7.52 con desviación estándar de 0.024. El análisis de 5 muestras de *B* mostró pH 7.49 con desviación estándar de 0.032. Con un nivel de significancia de 0.05 determine si los dos tipos de solución tienen valores diferentes de pH.

7.83. En una evaluación de psicología, 12 estudiantes de una clase obtuvieron una nota promedio de 78, con desviación estándar de 6, mientras que 15 estudiantes de otra clase tuvieron una nota promedio de 74, con desviación estándar de 8. Con un nivel de significancia de 0.05 determine si el primer grupo es superior al segundo.

PRUEBAS QUE INVOLUCRAN LA DISTRIBUCIÓN CHI CUADRADO

7.84. Se reporta que la desviación estándar de la resistencia al rompimiento de ciertos cables producidos por una compañía es 240 lb. Después de que se introdujo un cambio en el proceso de producción de estos cables, la resistencia al rompimiento de una muestra de 8 cables mostró una desviación estándar de 300 lb. Investigue la significancia del aumento aparente en la variación usando un nivel de significancia de (*a*) 0.05, (*b*) 0.01.

7.85. La temperatura anual de una ciudad se obtiene calculando la media de la temperatura media del día 15 de cada mes. La desviación estándar de la temperatura anual de la ciudad a lo largo de un periodo de 100 años fue 16 grados centígrados. Se calculó que durante los pasados 15 años la desviación estándar de la temperatura anual fue 10 grados centígrados. Pruebe la hipótesis de que la temperatura en la ciudad se ha vuelto menos variable que en el pasado, usando un nivel de significancia de (*a*) 0.05, (*b*) 0.01.

7.86. En el problema 7.77 la muestra de 20 bombillas eléctricas reveló una desviación estándar en los tiempos de vida de 150 horas. ¿Concluiría usted que esto es inusual? Explique.

PRUEBAS QUE INVOLUCRAN LA DISTRIBUCIÓN *F*

7.87. Dos muestras de 21 y 9 observaciones tienen varianzas dadas por $s_1^2 = 16$ y $s_2^2 = 8$, respectivamente. Pruebe la hipótesis de que la varianza de la primera población es mayor que la de la segunda con un nivel de significancia de (*a*) 0.05, (*b*) 0.01.

7.88. Desarrolle el problema 7.87 si las dos muestras constan de 60 y 120 observaciones, respectivamente.

7.89. ¿Podemos concluir en el problema 7.82 que hay una diferencia significativa en la variabilidad de los valores de pH de las dos soluciones al nivel de significancia de 0.10?

CURVAS DE OPERACIÓN CARACTERÍSTICAS

7.90. En relación con el problema 7.63 determine la probabilidad de aceptar la hipótesis de que hay proporciones iguales de canicas rojas y azules cuando la proporción real p de canicas rojas es (*a*) 0.6, (*b*) 0.7, (*c*) 0.8, (*d*) 0.9, (*e*) 0.3.

7.91. Represente los resultados del problema 7.90 en una gráfica de (*a*) β contra p, (*b*) $(1 - \beta)$ contra p. Compare las gráficas con las del problema 7.25, considerando la analogía de canicas rojas y azules a caras y sellos, respectivamente.

GRÁFICOS DE CONTROL DE CALIDAD

7.92. En el pasado cierto tipo de hilo producido por un fabricante tenía una resistencia promedio al rompimiento de 8.64 onzas y desviación estándar de 1.28 oz. Para determinar si la producción está conforme a los patrones, cada tres horas se toma una muestra de 16 trozos de hilo y se determina la resistencia promedio al rompimiento. Encuentre los límites de control del (*a*) 99.73% o 3σ, (*b*) 99% y (*c*) 95% en un gráfico de control de calidad y explique su aplicación.

7.93. En promedio cerca del 3% de los tornillos producidos por una fábrica son defectuosos. Para mantener este nivel de calidad se examina cada 4 horas una muestra de 200 tornillos. Determine los límites de control del (*a*) 99%, (*b*) 95% para el número de tornillos defectuosos en cada muestra. Observe que, en este caso, sólo se necesitan los *límites superiores de control*.

AJUSTE DE DATOS POR DISTRIBUCIONES TEÓRICAS

7.94. Ajuste una distribución binomial a los datos de la tabla 7-24.

Tabla 7-24

x	0	1	2	3	4
f	30	62	46	10	2

7.95. Ajuste una distribución normal a los datos del problema 5.98.

7.96. Ajuste una distribución normal a los datos del problema 5.100.

7.97. Ajuste una distribución de Poisson a los datos del problema 7.44 y compárela con el ajuste obtenido al usar la distribución binomial.

7.98. La tabla 7-25 muestra el número de muertes por año ocasionadas por patadas de caballo en 10 unidades del ejército prusiano, durante 20 años, desde 1875 a 1894. Ajuste una distribución de Poisson a los datos.

Tabla 7-25

x	0	1	2	3	4
f	109	65	22	3	1

LA PRUEBA CHI CUADRADO

7.99. En 60 lanzamientos de una moneda se observaron 37 caras y 23 sellos. Pruebe la hipótesis de que la moneda es balanceada, usando un nivel de significancia de (*a*) 0.05, (*b*) 0.01.

7.100. Desarrolle el problema 7.99 usando la corrección de Yates.

7.101. En un periodo de tiempo largo las notas dadas por un grupo de instructores en un curso en particular tuvieron en promedio: A 12%, B 18%, C 40%, D 18% y F 12%. Un instructor nuevo da 22 aes, 34 bes, 66 ces, 16 des y 12 efes durante dos semestres. Determine con un nivel de significancia de 0.05 si el instructor nuevo está siguiendo el patrón de notas de los otros instructores.

7.102. Se lanzan tres monedas juntas 240 veces, y cada vez se anota el número de caras que aparecen. Los resultados se muestran en la tabla 7-26 junto con los resultados esperados bajo la hipótesis de que las monedas son balanceadas. Pruebe la hipótesis con un nivel de significancia de 0.05.

Tabla 7-26

	0 caras	1 cara	2 caras	3 caras
Frecuencia observada	**24**	**108**	**95**	**23**
Frecuencia esperada	**30**	**90**	**90**	**30**

7.103. En la tabla 7-27 aparece el número de libros que prestó una biblioteca pública durante cierta semana. Pruebe la hipótesis de que el número de libros prestados no depende del día de la semana, con un nivel de significancia de (*a*) 0.05, (*b*) 0.01).

Tabla 7-27

	Lunes	Martes	Miércoles	Jueves	Viernes
Número de libros prestados	**135**	**108**	**120**	**114**	**146**

7.104. En una urna hay 6 canicas rojas y 3 blancas. De la urna se escogen al azar 2 canicas, se anota su color y se devuelven a la urna. Este proceso se llevó a cabo 120 veces y los resultados se anotaron en la tabla 7-28. (*a*) Determine las frecuencias esperadas. (*b*) Determine si los resultados obtenidos son consistentes con los esperados al nivel de significancia de 0.05.

Tabla 7-28

	0 rojas, 2 blancas	1 roja, 1 blanca	2 rojas, 0 blancas
Número de retiros	**6**	**53**	**61**

7.105. Se seleccionaron al azar 200 tornillos de la producción de 4 máquinas. Se encontró que el número de tornillos defectuosos fue 2, 9, 10, 3. Determine si existe una diferencia significativa entre las máquinas, usando un nivel de significancia de 0.05.

BONDAD DE AJUSTE

7.106. (*a*) Use la prueba chi cuadrado para determinar la bondad de ajuste de los datos en el problema 7.94. (*b*) ¿Es el ajuste "demasiado" bueno? Use un nivel de significancia de 0.05.

7.107. (*a*) Use la prueba chi cuadrado para determinar la bondad de ajuste de los datos en el (*a*) problema 7.95, (*b*) problema 7.96. Use un nivel de significancia de 0.05 y determine en cada caso si el ajuste es "demasiado" bueno.

7.108. Use la prueba chi cuadrado para determinar la bondad de ajuste de los datos en el (*a*) problema 7.97, (*b*) problema 7.98. ¿Es su resultado en (*a*) consistente con el del problema 7.106?

TABLAS DE CONTINGENCIA

7.109. La tabla 7-29 muestra el resultado de un experimento que investiga el efecto de la vacunación en animales de laboratorio contra cierta enfermedad. Usando un nivel de significancia de (*a*) 0.01, (*b*) 0.05 pruebe la hipótesis de que no hay diferencia entre los grupos vacunados y los no vacunados, es decir, la vacunación y esta enfermedad son independientes.

Tabla 7-29

	Se enfermaron	No se enfermaron
Vacunados	9	42
No vacunados	17	28

7.110. Desarrolle el problema 7.109 usando la corrección de Yates.

7.111. La tabla 7-30 muestra el número de estudiantes de dos clases *A* y *B*, que pasaron y perdieron un examen. Usando un nivel de significancia de (*a*) 0.05, (*b*) 0.01 pruebe la hipótesis de que no hay diferencia entre los dos cursos. Desarrolle el problema con y sin la corrección de Yates.

Tabla 7-30

	Pasaron	Perdieron
Clase A	72	17
Clase B	64	23

7.112. Algunas personas pertenecientes a un grupo de pacientes que se quejaban por no poder dormir bien recibieron pastillas para dormir, mientras que otras, del mismo grupo, recibieron pastillas de azúcar (y *creyeron* que estaban recibiendo pastillas para dormir). Luego, se les preguntó si las pastillas les habían ayudado o no. El resultado de sus respuestas aparece en la tabla 7-31. Suponiendo que todos los pacientes dijeron la verdad, pruebe la hipótesis de que no hay diferencia entre las pastillas para dormir y las pastillas de azúcar, con un nivel de significancia de 0.05.

Tabla 7-31

	Durmieron bien	No durmieron bien
Tomaron pastillas para dormir	44	10
Tomaron pastillas de azúcar	81	35

7.113. En una propuesta nacional de importancia, los miembros del partido A y del partido B votaron como aparece en la tabla 7-32. Con un nivel de significancia de (*a*) 0.01, (*b*) 0.05 pruebe la hipótesis de que no hay diferencia entre ambos partidos, en lo que a esta propuesta se refiere.

Tabla 7-32

	A favor	En contra	Indeciso
Partido A	85	78	37
Partido B	118	61	25

7.114. La tabla 7-33 muestra la relación entre el desempeño de los estudiantes en física y matemáticas. Pruebe la hipótesis de que el desempeño en física es independiente del desempeño en matemáticas, usando un nivel de significancia de (*a*) 0.05, (*b*) 0.01.

Tabla 7-33

		MATEMÁTICAS		
		Notas altas	Notas medias	Notas bajas
FÍSICA	Notas altas	56	71	12
	Notas medias	47	163	38
	Notas bajas	14	42	85

7.115. En la tabla 7-34 aparecen los resultado de una investigación realizada con el fin de determinar si la edad del conductor, de 21 años o mayor, tiene algún efecto en el número de accidentes en los que él está involucrado (incluyendo los accidentes menores). Con un nivel de significancia de (*a*) 0.05, (*b*) 0.01 pruebe la hipótesis de que el número de accidentes es independiente de la edad del conductor. ¿Qué fuentes de dificultad en la técnica de muestreo, así como otras consideraciones, pueden afectar sus conclusiones?

Tabla 7-34

		EDAD DEL CONDUCTOR				
		21–30	31–40	41–50	51–60	61–70
NÚMERO DE ACCIDENTES	0	748	821	786	720	672
	1	74	60	51	66	50
	2	31	25	22	16	15
	Más de 2	9	10	6	5	7

COEFICIENTES DE CONTINGENCIA

7.116. La tabla 7-35 muestra la relación entre el color del pelo y de los ojos de una muestra de 200 mujeres. (*a*) Encuentre el coeficiente de contingencia con y sin la corrección de Yates. (*b*) Compare el resultado de (*a*) con el coeficiente máximo de contingencia.

Tabla 7-35

		COLOR DE PELO	
		Rubio	No rubio
COLOR DE LOS OJOS	Azules	49	25
	No azules	30	96

7.117. Encuentre el coeficiente de contingencia para los datos de (*a*) el problema 7.109, (*b*) el problema 7.111 sin y con la corrección de Yates.

7.118. Encuentre el coeficiente de contingencia para los datos del problema 7.114.

PROBLEMAS DIVERSOS

7.119. Dos urnas, *A* y *B*, contienen igual número de canicas, pero las proporciones de canicas rojas y blancas en cada una es desconocida. La muestra de 50 canicas con remplazo de cada urna reveló 32 canicas rojas de la urna *A* y 23 de la urna *B*. Usando un nivel de significancia de 0.05 pruebe la hipótesis de que (*a*) las dos urnas tienen proporciones iguales de canicas y (*b*) *A* tiene mayor proporción de canicas rojas que *B*.

7.120. En relación con el problema 7.54 encuentre el menor número de preguntas que un estudiante debe contestar correctamente antes de que el instructor esté seguro de que el estudiante no está adivinando, con un nivel de significancia de (*a*) 0.05, (*b*) 0.01, (*c*) 0.06. Discuta los resultados.

7.121. Una moneda que se lanza 8 veces muestra 7 caras. ¿Podemos rechazar la hipótesis de que la moneda es balanceada con un nivel de significancia de (*a*) 0.05, (*b*) 0.10, (*c*) 0.01? Use una prueba de dos colas.

7.122. El porcentaje de *A* dadas en un curso de física en cierta universidad, en un periodo de tiempo largo, fue 10%. Durante un semestre en particular hubo 40 *A* en un grupo de 300 estudiantes. Pruebe la significancia de este resultado al nivel de (*a*) 0.05, (*b*) 0.01.

7.123. Usando gasolina de marca *A*, el número promedio de millas por galón que recorrieron 5 automóviles similares bajo condiciones idénticas fue 22.6, con desviación estándar de 0.48. Con gasolina de marca *B*, el número promedio fue 21.4, con desviación estándar 0.54. Con un nivel de significancia de 0.05, investigue si la marca *A* es realmente mejor que la marca *B* dando más recorrido por cada galón.

7.124. En el problema 7.123, ¿hay mayor variabilidad en millas por galón usando la marca *B* que usando la marca *A*? Explique.

Respuestas a los problemas suplementarios

7.63. (*a*) 0.2606

7.64. (*a*) Aceptar la hipótesis si se sacan entre 22 y 42 canicas rojas; de otra manera rechazarla. (*b*) 0.99. (*c*) Aceptar la hipótesis si se sacan entre 24 y 40 canicas rojas; de otra manera rechazarla.

7.65. (*a*) (H_0: $p = 0.5$) (H_1: $p > 0.5$). (*b*) Prueba de una cola. (*c*) Rechazar H_0 si se sacan más de 39 canicas rojas y aceptarla de otra manera (o abstenerse de tomar una decisión). (*d*) Rechazar H_0 si se sacan más de 41 canicas rojas y aceptarla de otra manera (o abstenerse de tomar una decisión).

7.66. (*a*) No podemos rechazar la hipótesis al nivel de 0.05.
(*b*) Podemos rechazar la hipótesis al nivel de 0.05.

7.67. No podemos rechazar la hipótesis al nivel de 0.01 ni en (*a*) ni en (*b*).

7.68. Podemos rechazar la afirmación a ambos niveles usando una prueba de una cola.

7.69. Sí, a ambos niveles, usando una prueba de una cola en cada caso.

7.70. El resultado es significativo al nivel de 0.05 tanto en la prueba de una cola como en la de dos colas.

7.71. El resultado es significativo al nivel de 0.01 en la prueba de una cola, pero no en la de dos colas.

7.72. (*a*) Sí. (*b*) No.

7.73. Una prueba de una cola muestra en ambos niveles de significancia que la marca *B* es superior a la marca *A*.

7.74. Una prueba de una cola muestra que la diferencia es significativa al nivel de 0.05, pero no al de 0.01.

7.75. Una prueba de una cola muestra que el fertilizante nuevo es superior en ambos niveles de significancia.

7.76. (*a*) Una prueba de dos colas muestra que no hay diferencia en la calidad del desempeño al nivel de 0.05.
(*b*) Una prueba de una cola muestra que *B* no se está desempeñando mejor que *A* al nivel de 0.05.

7.77. Una prueba de dos colas muestra que no hay evidencia en ninguno de los niveles de que el tiempo de vida promedio haya cambiado.

7.78. Una prueba de una cola indica que no hay disminución en la media ni al nivel de 0.05 ni al de 0.01.

7.79. Una prueba de dos colas a ambos niveles muestra que el producto no cumple con las especificaciones.

7.80. Una prueba de una cola a ambos niveles muestra que el contenido promedio de cobre es mayor que el requerido en las especificaciones.

7.81. Una prueba de una cola muestra que el proceso no se debe introducir si el nivel de significancia adoptado es 0.01, pero debe introducirse si el nivel de significancia adoptado es 0.05.

7.82. Usando una prueba de dos colas con nivel de significancia de 0.05, no podemos concluir que haya diferencia en la acidez.

7.83. Usando una prueba de una cola, con nivel de significancia de 0.05, podemos concluir que el primer grupo no es superior al segundo.

7.84. El aparente aumento en la variabilidad no es significativo en ambos niveles.

7.85. La aparente disminución es significativa al nivel de 0.05, pero no al de 0.01.

7.86. Podemos concluir que el resultado es inusual al nivel de 0.05, pero no al de 0.01.

7.87. No podemos concluir que la primera varianza sea mayor que la segunda en cualquier nivel.

7.88. Podemos concluir que la primera varianza es mayor que la segunda en ambos niveles.

7.89. No.

7.90. (*a*) 0.311 (*b*) 0.0118 (*c*) 0 (*d*) 0 (*e*) 0.0118.

7.92. (*a*) 8.64 ± 0.96 oz (*b*) 8.64 ± 0.83 oz (*c*) 8.64 ± 0.63 oz

7.93. (*a*) 6 (*b*) 4

7.94. $f(x) = {}_4C_x(0.32)^x(0.68)^{4-x}$; las frecuencias esperadas son 32, 60, 43, 13 y 2, respectivamente.

7.95. Las frecuencias esperadas son 1.7, 5.5, 12.0, 15.9, 13.7, 7.6, 2.7 y 0.6, respectivamente.

7.96. Las frecuencias esperadas son 1.1; 4.0, 11.1, 23.9, 39.5, 50.2, 49.0, 36.6, 21.1, 9.4, 3.1 y 1.0, respectivamente.

7.97. Las frecuencias esperadas son 41.7, 53.4, 34.2, 14.6 y 4.7, respectivamente.

7.98. $f(x) = \dfrac{(0.61)^x e^{-0.61}}{x!}$; las frecuencias esperadas son 108.6, 66.3, 20.2, 4.1 y 0.7, respectivamente.

7.99. La hipótesis no puede rechazarse a cualquier nivel.

7.100. La conclusión es la misma que la anterior.

7.101. El instructor nuevo no está siguiendo el patrón de calificación de los otros (el hecho de que las notas parezcan ser mejores que el promedio *puede* deberse a una mejor habilidad en la enseñanza o a patrones más bajos o ambos).

7.102. No hay razón para rechazar la hipótesis de que las monedas sean balanceadas.

7.103. No hay razón para rechazar la hipótesis en cualquier nivel.

7.104. (*a*) 10, 60, 50, respectivamente. (*b*) La hipótesis de que los resultados son los mismos a los esperados no puede rechazarse al nivel de significancia de 0.05.

7.105. La diferencia es significativa al nivel de 0.05.

7.106. (*a*) El ajuste es bueno. (*b*) No.

7.107. (*a*) El ajuste es "demasiado" bueno. (*b*) El ajuste es pobre al nivel de 0.05.

7.108. (*a*) El ajuste es muy pobre al nivel de 0.05. Ya que la distribución binomial da un buen ajuste de los datos, esto es consistente con el problema 7.109. (*b*) El ajuste es bueno, pero no "muy bueno".

7.109. La hipótesis puede rechazarse al nivel de 0.05, pero no al nivel de 0.01.

7.110. La misma conclusión.

7.111. La hipótesis no puede rechazarse a cualquier nivel.

7.112. La hipótesis no puede rechazarse al nivel de 0.05.

7.113. La hipótesis puede rechazarse en ambos niveles.

7.114. La hipótesis puede rechazarse en ambos niveles.

7.115. La hipótesis no puede rechazarse a cualquier nivel.

7.116. (*a*) 0.3863, 0.3779 (con la corrección de Yates)

7.117. (*a*) 0.2205, 0.1985 (corregido) (*b*) 0.0872, 0.0738 (corregido)

7.118. 0.4651

7.119. (*a*) Una prueba de dos colas al nivel de 0.05 fracasa al rechazar la hipótesis de proporciones iguales.

(*b*) Una prueba de una cola al nivel de 0.05 indica que *A* tiene mayor proporción de canicas rojas que *B*.

7.120. (*a*) 9 (*b*) 10 (*c*) 10 (*d*) 8.

7.121. (*a*) No (*b*) Sí (*c*) No.

7.122. Usando una prueba de una cola el resultado es significativo al nivel de 0.05, pero no al nivel de 0.01.

7.123. Al nivel de 0.05 podemos concluir que la marca *A* es mejor que la marca *B*.

7.124. No al nivel de 0.05.

CAPÍTULO 8

Curva de ajuste, regresión y correlación

CURVA DE AJUSTE

Con frecuencia, en la práctica se encuentra que existe una relación entre dos (o más variables y uno desea expresar esta relación de manera matemática planteando una ecuación que conecte las variables.

Un primer paso es recolectar los datos mostrando los valores correspondientes de las variables. Por ejemplo, supongamos que x y y denotan, respectivamente, la estatura y el peso de un varón adulto. Entonces, una muestra de n individuos revelaría las estaturas $x_1, x_2, \ldots, x_n$ y los pesos correspondientes $y_1, y_2, \ldots, y_n$.

El siguiente paso es dibujar los puntos $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ en un sistema rectangular de coordenadas. El conjunto resultante de puntos se denomina *diagrama de dispersión*.

A partir del diagrama de dispersión a menudo es posible visualizar una curva suave que se aproxima a los datos. Tal curva se conoce como *curva de aproximación*. En la figura 8-1, por ejemplo, los datos se aproximan bastante bien por una línea recta, y decimos entonces que existe una *relación lineal* entre las variables. Sin embargo, aunque en la figura 8-2 existe una relación entre las variables, ésta no es lineal, así que la llamamos *relación no lineal*. En la figura 8-3 parece no haber relación entre las variables.

El problema general de encontrar ecuaciones de curvas de aproximación que se ajusten a conjuntos dados de datos se llama *curva de ajuste*. En la práctica el tipo de ecuación se sugiere a partir del diagrama de dispersión. Para la figura 8-1 podemos usar una línea recta

$$y = a + bx \tag{1}$$

mientras que para la figura 8-2 podemos usar la *parábola* o *curva cuadrática*.

$$y = a + bx + cx^2 \tag{2}$$

Algunas veces ayuda dibujar diagramas de dispersión en términos de *variables transformadas*. Por ejemplo, si log y contra x lleva a una línea recta, podemos tratar $\log y = a + bx$ como ecuación para la curva de aproximación.

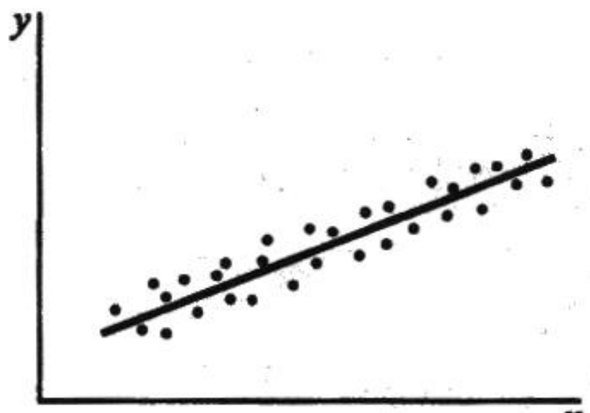


Fig. 8-1

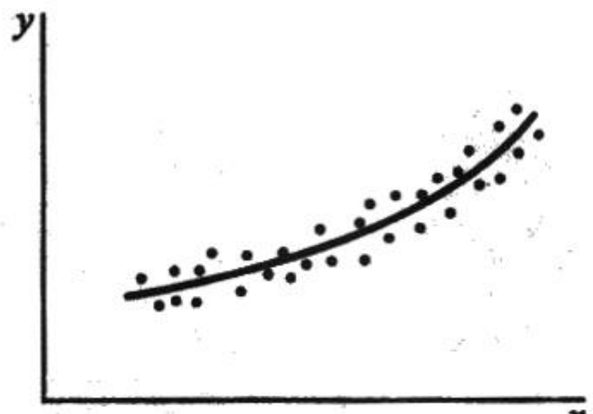


Fig. 8-2

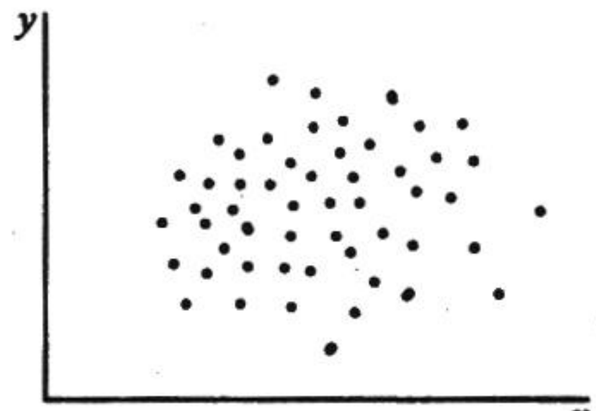


Fig. 8-3

REGRESIÓN

Uno de los principales propósitos de la curva de ajuste es estimar una de las variables (*la variable dependiente*) a partir de la otra (*la variable independiente*). El proceso de estimación se conoce como *regresión*. Si y va a estimarse a partir de x con ayuda de alguna ecuación, denominamos dicha ecuación una *ecuación de regresión de y en x* y la curva correspondiente, *curva de regresión de y en x*.

MÉTODO DE LOS MÍNIMOS CUADRADOS

Por lo general, más de una curva de un tipo dado parece ajustarse al conjunto de datos. Para evitar el juicio individual en la construcción de rectas, parábolas u otras curvas de aproximación, es necesario estar de acuerdo con la definición de "mejor línea de ajuste", "mejor parábola de ajuste", etc.

Para obtener una definición posible, consideremos la figura 8-4, en donde los puntos de los datos son $(x_1, y_1), \ldots, (x_n, y_n)$. Para un valor dado de x, digamos x_1, habrá una diferencia entre el valor y_1 y el valor correspondiente, como se determina en la curva C. Denotamos esta diferencia como d_1, la cual denominamos desviación, error o residuo, y que puede ser positiva, negativa o cero. De igual manera, a los valores $x_2, \ldots, x_n$, les corresponden las desviaciones $d_2, \ldots d_n$.

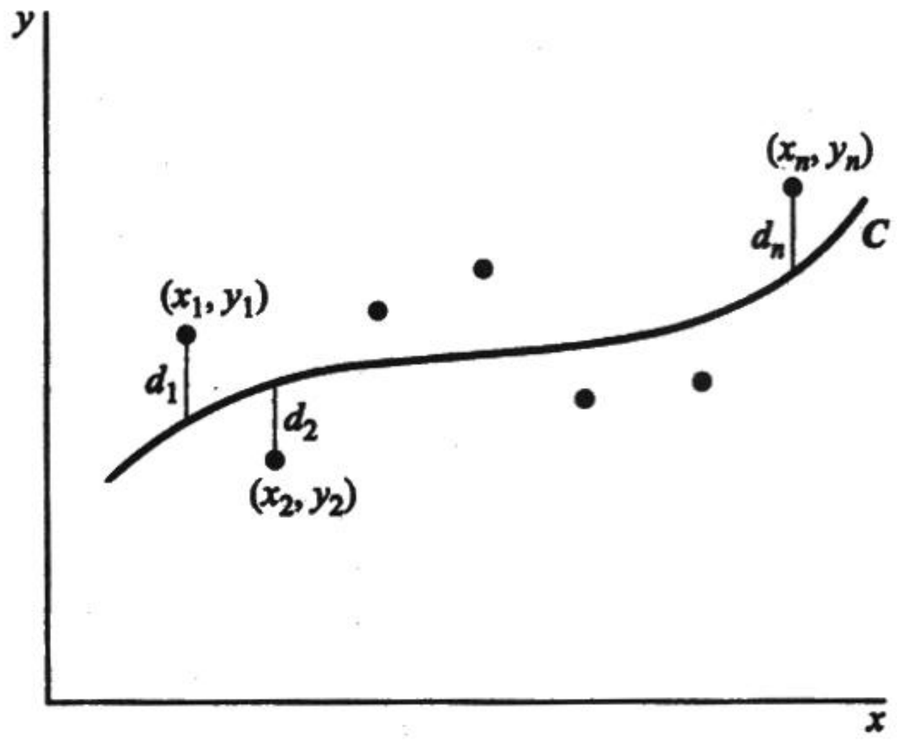


Fig. 8-4

La cantidad $d_1^2 + d_2^2 + \cdots + d_n^2$ provee una medida de la bondad de ajuste de la curva C. Si ésta es pequeña, el ajuste es bueno, si es grande, el ajuste es malo. Por consiguiente, ahora podemos plantear la siguiente definición.

Definición. De todas las curvas de una familia dada de curvas que se aproximan al conjunto de n puntos de datos, la curva que tenga la propiedad de

$$d_1^2 + d_2^2 + \cdots + d_n^2 = \text{un mínimo}$$

se llama la *mejor curva de ajuste* de la familia.

Se dice que una curva que tenga esta propiedad ajusta los datos en el *sentido de mínimos cuadrados*, y se llama *curva de regresión de mínimos cuadrados*, o simplemente, *curva de mínimos cuadrados*. Una línea con esta propiedad recibe el nombre de *recta de mínimos cuadrados*; una parábola con esta propiedad se llama *parábola de mínimos cuadrados*, etc.

Se acostumbra usar la definición anterior cuando x es la variable independiente y y la variable dependiente. Si x es la variable dependiente, la definición se modifica considerando variaciones horizontales en lugar de variaciones verticales, lo cual significa intercambiar los ejes x y y. Estas dos definiciones llevan, en general, a dos curvas diferentes de mínimos cuadrados. A menos que se especifique otra cosa, debemos considerar y como la variable dependiente y x como la variable independiente.

Es posible definir otra curva de mínimos cuadrados considerando distancias perpendiculares desde los puntos de datos a la curva en lugar de distancias verticales y horizontales. Sin embargo, esto no se usa con frecuencia.

RECTA DE MÍNIMOS CUADRADOS

Usando la definición anterior podemos demostrar (*véase* problema 8.3) que la recta de mínimos cuadrados que se aproxima al conjunto de puntos $(x_1, y_1), \ldots, (x_n, y_n)$ tiene la siguiente ecuación

$$y = a + bx \tag{3}$$

donde las constantes a y b se determinan resolviendo simultáneamente las ecuaciones

$$\begin{aligned} \sum y &= an + b \sum x \\ \sum xy &= a \sum x + b \sum x^2 \end{aligned} \tag{4}$$

que son las *ecuaciones normales* para la recta de mínimos cuadrados. Observe que por brevedad usamos $\sum y$, $\sum xy$ en lugar de $\sum_{j=1}^{n} y_j$, $\sum_{j=1}^{n} x_j y_j$. Las ecuaciones normales (*4*) se pueden recordar con facilidad si tenemos en cuenta que la primera ecuación se puede obtener formalmente sumando ambos lados de (*3*), mientras que la segunda se obtiene formalmente primero multiplicando ambos lados de (*3*) por x y luego sumando. Claro que esto no es una derivación de las ecuaciones normales, sino sólo una forma de recordarlas.

Los valores a y b obtenidos de (*4*) están dados por

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n \sum x^2 - (\sum x)^2} \qquad b = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2} \tag{5}$$

El resultado para b en (*5*) también se puede escribir

$$b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} \tag{6}$$

Como es usual, aquí una barra indica *promedio* o *media*, así $\bar{x} = (\sum x)/n$. De la división de ambos lados de la primera ecuación normal en (*4*) por n se obtiene

$$\bar{y} = a + b\bar{x} \tag{7}$$

Si se quiere, podemos encontrar primero b a partir de (*5*) o (*6*) y luego usar (*7*) para encontrar $a = \bar{y} - b\bar{x}$. Esto equivale a escribir la recta de mínimos cuadrados como

$$y - \bar{y} = b(x - \bar{x}) \qquad \text{o} \qquad y - \bar{y} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} (x - \bar{x}) \tag{8}$$

El resultado (*8*) muestra que la constante b, que es la inclinación o *pendiente* de la curva (*3*), es la constante fundamental para determinar la recta. A partir de (*8*) también se puede ver que la recta de mínimos cuadrados pasa a través del punto $(\bar{x}, \bar{y})$, el cual se denomina *centroide* o *centro de gravedad de los datos*.

La pendiente b de la recta de regresión es independiente del origen de las coordenadas. Esto significa que si hacemos la transformación (llamada *traslación de ejes*) dada por

$$x = x' + h \qquad y = y' + k \tag{9}$$

donde h y k son constantes cualesquiera, entonces b también está dada por

$$b = \frac{n\sum x'y' - (\sum x')(\sum y')}{n\sum x'^2 - (\sum x')^2} = \frac{\sum(x' - \overline{x'})(y' - \overline{y'})}{\sum(x' - \overline{x'})^2} \tag{10}$$

donde x, y se remplazaron por x', y' [por esta razón decimos que b es *invariante bajo la transformación* (*9*)]. También se debe notar, sin embargo, que a, que determina el intercepto en el eje x, depende del origen (y por lo tanto no es invariante).

En el caso particular donde $h = \bar{x}$, $k = \bar{y}$, (*10*) se simplifica a

$$b = \frac{\sum x'y'}{\sum x'^2} \tag{11}$$

Los resultados (*10*) u (*11*) se usan para simplificar el trabajo requerido para obtener la recta de mínimos cuadrados.

Las observaciones anteriores también son válidas para la curva de regresión de x en y. Los resultados se obtienen simplemente intercambiando x y y. Por ejemplo, la recta de regresión de mínimos cuadrados de x en y es

$$x - \bar{x} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(y - \bar{y})^2}(y - \bar{y}) \tag{12}$$

Se debe notar que en general (*12*) no es la misma recta que (*8*).

RECTA DE MÍNIMOS CUADRADOS EN TÉRMINOS DE VARIANZA MUESTRAL Y COVARIANZA

Las varianzas muestrales y la covarianza de x y y están dadas por

$$s_x^2 = \frac{\sum(x - \bar{x})^2}{n}, \qquad s_y^2 = \frac{\sum(y - \bar{y})^2}{n}, \qquad s_{xy} = \frac{\sum(x - \bar{x})(y - \bar{y})}{n} \tag{13}$$

En términos de éstas, las rectas de regresión de mínimos cuadrados de y en x y de x en y pueden escribirse, respectivamente, como

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x}) \qquad \text{y} \qquad x - \bar{x} = \frac{s_{xy}}{s_y^2}(y - \bar{y}) \tag{14}$$

si definimos formalmente el *coeficiente de correlación muestral* por [compare con (*54*), página 89]

$$r = \frac{s_{xy}}{s_x s_y} \tag{15}$$

entonces (*14*) se puede escribir como

$$\frac{y - \bar{y}}{s_y} = r\left(\frac{x - \bar{x}}{s_x}\right) \qquad \text{y} \qquad \frac{x - \bar{x}}{s_x} = r\left(\frac{y - \bar{y}}{s_y}\right) \tag{16}$$

En vista de que $(x - \bar{x})/S_x$ y $(y - \bar{y})/S_y$ son valores muestrales estandarizados o valores estándar, los resultados en (*16*) nos ofrecen una manera muy sencilla para recordar las rectas de regresión. Es claro que las dos rectas en (*16*) son diferentes a menos que $r = \pm 1$, en cuyo caso todos los puntos muestrales caen en una recta [esto se mostrará en (*26*)] y hay una *correlación y regresión lineal perfecta*.

Es también interesante anotar que si las dos rectas de regresión (16) se escriben como $y = a + bx$, $x = c + dy$, respectivamente, entonces

$$bd = r^2 \tag{17}$$

Hasta el momento no hemos considerado la significancia precisa del coeficiente de correlación, sino que sólo lo definimos formalmente en términos de varianzas y covarianza. En la página 290, daremos la significancia.

PARÁBOLA DE MÍNIMOS CUADRADOS

Los conceptos anteriores pueden ampliarse fácilmente. Por ejemplo, la *parábola de mínimos cuadrados* que ajusta un conjunto de puntos muestrales está dada por

$$y = a + bx + cx^2 \tag{18}$$

donde a, b, c se determinan a partir de las ecuaciones normales

$$\begin{aligned} \sum y &= na + b\sum x + c\sum x^2 \\ \sum xy &= a\sum x + b\sum x^2 + c\sum x^3 \\ \sum x^2y &= a\sum x^2 + b\sum x^3 + c\sum x^4 \end{aligned} \tag{19}$$

Éstas se obtienen sumando ambos lados de (*18*) después de multiplicar sucesivamente por 1, x y x^2, respectivamente.

REGRESIÓN MÚLTIPLE

Los conceptos anteriores pueden generalizarse a más variables. Por ejemplo, si creemos que hay una relación lineal entre la variable dependiente z y dos variables independientes x y y, entonces podemos buscar una ecuación que conecte las variables y tenga la forma

$$z = a + bx + cy \tag{20}$$

Ésta se llama una *ecuación de regresión de z en x* y *en y*. Si x es la variable dependiente, entonces una ecuación similar se llamaría *ecuación de regresión de x en y* y *en z*.

Debido a que (*20*) representa un plano en un sistema de coordenada rectangular de tres dimensiones, se denomina *plano de regresión*. Para encontrar el plano de regresión de mínimos cuadrados, determinamos a, b, c en (*20*) de manera que

$$\begin{aligned} \sum z &= na + b\sum x + c\sum y \\ \sum xz &= a\sum x + b\sum x^2 + c\sum xy \\ \sum yz &= a\sum y + b\sum xy + c\sum y^2 \end{aligned} \tag{21}$$

Estas ecuaciones, llamadas *ecuaciones normales* que corresponden a (*20*), se obtienen como resultado de aplicar una definición similar a la de la página 287. Observe que se pueden obtener formalmente de (*20*) multiplicando por 1, x, y, respectivamente, y sumando.

Se pueden hacer fácilmente generalizaciones a más variables, involucrando ecuaciones lineales y no lineales que llevan a *superficies de regresión* en espacios de tres o más dimensiones.

ERROR ESTÁNDAR DEL ESTIMADO

Si y_{est} es el valor estimado de y para un valor dado de x, como se obtiene de la curva de regresión de y en x, entonces la medida de la dispersión alrededor de la curva de regresión está dada por la cantidad

$$s_{y\cdot x} = \sqrt{\frac{\sum (y - y_{est})^2}{n}} \tag{22}$$

la cual se denomina *error estándar del estimado de y en x*. Dado que $\sum(y - y_{est})^2 = \sum d^2$, como se usó en la definición de la página 286, vemos que de todas las posibles curvas de regresión, la curva de mínimos cuadrados tiene el menor error estándar del estimado.

En el caso de una recta de regresión $y_{est} = a + bx$, con a y b dados en (*4*), tenemos

$$s_{y\cdot x}^2 = \frac{\sum y^2 - a\sum y - b\sum xy}{n} \tag{23}$$

o
$$s_{y\cdot x}^2 = \frac{\sum (y-\bar{y})^2 - b\sum (x-\bar{x})(y-\bar{y})}{n} \tag{24}$$

También podemos expresar $s_{y\cdot x}^2$ para la recta de mínimos cuadrados en términos de la varianza y el coeficiente de correlación como

$$s_{y\cdot x}^2 = s_y^2(1-r^2) \tag{25}$$

a partir de lo cual se deduce como corolario que $r^2 \leqq 1$, es decir, $-1 \leqq r \leqq 1$.

El error estándar del estimado tiene propiedades análogas a las de la desviación estándar. Por ejemplo, si construimos pares de líneas paralelas a la línea de regresión de y en x a distancias verticales respectivas $s_{y\cdot x}$, $2s_{y\cdot x}$ y $3s_{y\cdot x}$ de ella, podemos encontrar, si n es lo suficientemente grande como para que esté incluido entre esos pares de líneas, alrededor del 68%, 95% y 99.7% de los puntos muestrales, respectivamente (*véase* problema 8.23).

Así como hay un estimador sin sesgo de la varianza de la población dado por $\hat{s}^2 = ns^2/(n-1)$, así mismo hay un estimador insesgado del cuadrado del estimado del error estándar. Esto está dado por $\hat{s}_{y\cdot x}^2 = ns_{y\cdot x}^2/(n-2)$. Por esta razón algunos estadísticos prefieren expresar el denominador en (22) con $n-2$ en lugar de n.

Las observaciones anteriores se pueden modificar con facilidad para la recta de regresión de x en y (en cuyo caso el error estándar del estimador se denota por $s_{x\cdot y}$) o para la regresión no lineal o múltiple.

COEFICIENTE DE CORRELACIÓN LINEAL

Hasta ahora definimos el coeficiente de correlación según (*15*), pero no hemos examinado su importancia. Para hacerlo, observemos que a partir de (*25*) y de las definiciones de $s_{y\cdot x}$ y s_y, tenemos

$$r^2 = 1 - \frac{\sum (y-y_{\text{est}})^2}{\sum (y-\bar{y})^2} \tag{26}$$

Ahora podemos ver que (*véase* problema 8.24)

$$\sum (y-\bar{y})^2 = \sum (y-y_{\text{est}})^2 + \sum (y_{\text{est}}-\bar{y})^2 \tag{27}$$

La cantidad de la izquierda de (*27*) se llama *variación total*. La primera suma a la derecha de (*27*) se llama *variación no explicada*, mientras que la segunda suma se llama *variación explicada*. Esta terminología surge debido a que las desviaciones $y - y_{\text{est}}$ se comportan de manera aleatoria o impredecible, mientras que las desviaciones $y_{\text{est}} - \bar{y}$ se explican por la recta de regresión de mínimos cuadrados y tienden a seguir patrones definidos. Se deduce de (*26*) y (*27*) que

$$r^2 = \frac{\sum (y_{\text{est}}-\bar{y})^2}{\sum (y-\bar{y})^2} = \frac{\text{variación explicada}}{\text{variación total}} \tag{28}$$

Por consiguiente, r^2 se puede interpretar como la fracción de la variación total que se explica por la recta de regresión de mínimos cuadrados. En otras palabras, r mide *qué tan bien* se ajusta la recta de regresión de mínimos cuadrados a los datos muestrales. Si la variación total se explica *toda* por la recta de regresión, es decir, si $r^2 = 1$ o $r \pm 1$, decimos que hay una *correlación lineal perfecta* (y en tal caso *regresión lineal perfecta*). Por otro lado, si toda la variación total es no explicada, entonces la variación explicada es cero y así $r = 0$. En la práctica la cantidad r^2, que se denomina coeficiente de determinación, se ubica entre 0 y 1.

El coeficiente de correlación se puede calcular a partir de cualquiera de los resultados

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2}\sqrt{\sum (y-\bar{y})^2}} \tag{29}$$

o

$$r^2 = \frac{\text{variación explicada}}{\text{variación total}} = \frac{\sum (y_{est} - \bar{y})^2}{\sum (y - \bar{y})^2} \tag{30}$$

los cuales son equivalentes para la regresión lineal. La fórmula (*29*) se conoce como la *fórmula de producto-momento* para la correlación lineal.

Fórmulas equivalentes a las anteriores, que en la práctica se usan con frecuencia, son

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \tag{31}$$

y

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}} \tag{32}$$

Si usamos la transformación (*9*), página 287, encontramos

$$r = \frac{n\sum x'y' - (\sum x')(\sum y')}{\sqrt{[n\sum x'^2 - (\sum x')^2][n\sum y'^2 - (\sum y')^2]}} \tag{33}$$

que muestra que r es no variante bajo una traslación de ejes. En particular, si $h = \bar{x}$, $k = \bar{y}$, (*33*) se convierte en

$$r = \frac{\sum x'y'}{\sqrt{(\sum x'^2)(\sum y'^2)}} \tag{34}$$

que con frecuencia es útil en los cálculos.

El coeficiente de correlación lineal puede ser positivo o negativo. Si r es positivo, y tiende a aumentar cuando aumenta x (la pendiente de la recta de mínimos cuadrados es positiva); mientras que si r es negativo, y tiende a disminuir cuando aumenta x (la pendiente es negativa). El signo se toma *automáticamente* en cuenta si usamos el resultado (*29*), (*31*), (*32*), (*33*) o (*34*). Sin embargo, si usamos (*30*) para obtener r, debemos aplicar el signo adecuado.

COEFICIENTE DE CORRELACIÓN GENERALIZADO

La definición (*29*) [o alguna de sus formas equivalentes de (*31*) a (*34*)] para el coeficiente de correlación involucra sólo los valores muestrales x, y. En consecuencia, da el mismo número para todas las formas de curvas de regresión y es inútil en la medición del ajuste, excepto en el caso de la regresión lineal, donde da la casualidad de que coincide con (*30*). Sin embargo, la última definición, es decir,

$$r^2 = \frac{\text{variación explicada}}{\text{variación total}} = \frac{\sum (y_{est} - \bar{y})^2}{\sum (y - \bar{y})^2} \tag{35}$$

refleja la forma de la curva de regresión (via y_{est}) y así es apropiada como la definición del *coeficiente de correlación generalizado r*. Usamos (*35*) para obtener coeficientes de correlación no lineal (los cuales miden qué tan bien se ajustan los datos a una *curva de regresión no lineal*) o, por generalización apropiada, los *coeficientes de correlación múltiple*. La conexión (*25*) entre el coeficiente de correlación y el error estándar o estimado es válida también para la correlación no lineal.

Dado que el coeficiente de correlación simplemente mide qué tan bien una curva de regresión dada (o superficie) ajusta datos muestrales, es claro que no tiene sentido usar un coeficiente de correlación lineal donde los datos no son lineales. Supongamos, sin embargo, que uno aplica (*29*) a datos no lineales y se obtiene un valor que numéricamente es considerablemente menor que 1. Entonces, la conclusión que se saca no es que hay *poca correlación* (conclusión a la que llegan las personas poco familiarizadas con los fundamentos de la teoría de la correlación), sino que hay *poca correlación lineal*. De hecho, puede existir una *gran* correlación no lineal.

CORRELACIÓN DE RANGO

En lugar de usar valores muestrales precisos, o cuando la precisión es inalcanzable, los datos se pueden organizar en orden de tamaño, importancia, etc., usando los números 1, 2, . . ., n. Si dos conjuntos correspondientes de valores x y y se organizan de tal manera, *el coeficiente de correlación de rango*, que se denota como r_{rank}, o brevemente r, está dado por (*véase* el problema 8.36)

$$r_{\text{rank}} = 1 - \frac{6\sum d^2}{n(n^2-1)} \tag{36}$$

donde d = diferencias entre las clasificaciones de los correspondientes x y y
n = número de pares de valores (x, y) en los datos

La cantidad r_{rank} en (*36*) se conoce como el coeficiente de correlación de rango de Spearman.

INTERPRETACIÓN PROBABILÍSTICA DE LA REGRESIÓN

Un diagrama de dispersión, como el de la figura 8-1, es una representación gráfica de los puntos de datos para una muestra en particular. Si se escoge una muestra diferente o se agranda la muestra original, se obtiene, en general, un diagrama de dispersión algo diferente. Cada diagrama de dispersión da como resultado una recta o curva de regresión diferente, aunque esperamos que éstas no difieran significativamente entre sí, si las muestras se tomaron de la misma población.

A partir del concepto de curva de ajuste en muestras, llegamos a la curva de ajuste para la población de la cual se toman las muestras. La dispersión de los puntos alrededor de la recta o curva de regresión indica que para un valor particular de x, hay en realidad varios valores de y *distribuidos* alrededor de la recta o curva. Este concepto de distribución nos lleva naturalmente a darnos cuenta de que hay una conexión entre la curva de ajuste y la probabilidad.

La conexión se suministra al introducir las variables aleatorias X y Y, las cuales pueden asumir valores muestrales diferentes x y y, respectivamente. Por ejemplo X y Y pueden representar las estaturas y pesos de varones adultos en una población de donde se obtienen muestras. Entonces, suponemos que X y Y tienen una función de probabilidad conjunta, $f(x, y)$, según se consideren discretas o continuas.

Dada la función de densidad conjunta o función de probabilidad, $f(x, y)$, de dos variables aleatorias X y Y, es natural preguntarse si hay una función $g(X)$ tal que

$$E\{[Y - g(X)]^2\} = \text{un mínimo} \tag{37}$$

Una curva con la ecuación $y = g(x)$ que tiene la propiedad (*37*) se llama *curva de regresión de mínimos cuadrados de Y en X*. Tenemos el siguiente teorema:

***Teorema 8-1*:** Si X y Y son variables aleatorias con función de densidad conjunta o función de probabilidad $f(x, y)$, entonces existe una curva de regresión de mínimos cuadrados de Y en X con la propiedad (*37*) dada por

$$y = g(x) = E(Y \mid X = x) \tag{38}$$

siempre y cuando X y Y tengan una varianza finita.

Observe que $E(Y \mid X = x)$ es la esperanza condicional de Y dado $X = x$, como se definió en la página 89.

Se pueden hacer observaciones similares para la curva de regresión de mínimos cuadrados de X y Y. En ese caso (*37*) se remplaza por

$$E\{[X - h(Y)^2\} = \text{un mínimo}$$

y (*38*) se remplaza por $x = h(y) = E(X \mid Y = y)$. Las dos curvas de regresión $y = g(x)$ y $x = h(y)$ son diferentes en general.

Surge un caso interesante cuando la distribución conjunta es la distribución normal bidimensional dada por (*49*), página 127. Tenemos entonces el siguiente teorema:

Teorema 8-2: Si X y Y son variables aleatorias con distribución normal bidimensional, entonces la curva de regresión de mínimos cuadrados de Y en X es una recta de regresión dada por

$$\frac{y-\mu_Y}{\sigma_Y} = \rho\left(\frac{x-\mu_X}{\sigma_X}\right) \tag{39}$$

donde

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{40}$$

representa el coeficiente de correlación de la población.

También podemos escribir (*39*) como

$$y - \mu_Y = \beta(x - \mu_X) \tag{41}$$

donde

$$\beta = \frac{\sigma_{XY}}{\sigma_X^2} \tag{42}$$

Se pueden hacer observaciones similares para la curva de regresión de mínimos cuadrados de X en Y, que también es una recta [dado por (*39*) con X y Y, x y y, intercambiadas]. Estos resultados se deben comparar con los resultados correspondientes en la página 288.

En caso de que $f(x, y)$ sea desconocida, podemos incluso usar el criterio (*37*) para obtener curvas aproximadas de regresión para la población. Por ejemplo, si suponemos $g(x) = \alpha + \beta x$, obtenemos la recta de regresión de mínimos cuadrados (*39*), donde α, β están dadas en términos de los parámetros (desconocidos) μ_X, μ_Y, σ_X, σ_Y, ρ. De igual manera, si $g(x) = \alpha + \beta x + \gamma x^2$, podemos obtener una parábola de regresión de mínimos cuadrados, etc. *Véase* el problema 8.39.

En general, todas las observaciones hechas para muestras en las páginas 286 a 291, pueden extenderse fácilmente a la población. Por ejemplo, el error estándar o estimado en el caso de la población está dado en términos de la varianza y el coeficiente de correlación de

$$\sigma_{Y.X}^2 = \sigma_Y^2(1 - \rho^2) \tag{43}$$

el cual se debe comparar con (*25*), página 290.

INTERPRETACIÓN PROBABILÍSTICA DE LA CORRELACIÓN

A partir de las observaciones anteriores es claro que el coeficiente de correlación de la población debe dar una medida de qué tan bien se ajusta una curva de regresión de una población dada a los datos de la población. Todas las observaciones anteriores hechas a la correlación en una muestra aplican también a la población. Por ejemplo, si $g(x)$ se determina por (*37*), entonces

$$E[(Y - \bar{Y})^2] = E[(Y - Y_{\text{est}})^2] + E[(Y_{\text{est}} - \bar{Y})^2] \tag{44}$$

donde $Y_{\text{est}} = g(X)$ y $\bar{Y} = E(Y)$. Las tres cantidades en (*44*) reciben el nombre de *variación total, no explicada* y *explicada*, respectivamente. Esto nos lleva a la definición del *coeficiente de correlación de la población* ρ, donde

$$\rho^2 = \frac{\text{variación explicada}}{\text{variación total}} = \frac{E[(Y_{\text{est}} - \bar{Y})^2]}{E[(Y - \bar{Y})^2]} \tag{45}$$

Para el caso lineal esto se reduce a (*40*). También se pueden escribir resultados similares del (*31*) al (*34*) para el caso de una población y regresión lineal. El resultado (*45*) también se utiliza para definir ρ en el caso no lineal.

TEORÍA MUESTRAL DE LA REGRESIÓN

La ecuación de regresión $y = a + bx$ se obtiene con base en datos muestrales. Con frecuencia estamos interesados en la ecuación de regresión correspondiente $y = \alpha + \beta x$ para la población de la cual se tomó la muestra. Las siguientes son algunas pruebas concernientes a la población normal. Para mantener la notación de forma simple, debemos seguir la convención común de indicar los valores de las variables aleatorias muestrales, en lugar de las variables aleatorias mismas.

1. **PRUEBA DE HIPÓTESIS** $\beta = b$. Para probar la hipótesis de que el coeficiente de regresión β es igual a algún valor especificado de b, usamos el hecho de que el estadístico

$$t = \frac{\beta - b}{s_{y.x}/s_x}\sqrt{n-2} \tag{46}$$

tiene la distribución de Student con $n - 2$ grados de libertad. Esto puede usarse, también, para encontrar los intervalos de confianza para los coeficientes de regresión de la población a partir de valores muestrales. *Véanse* los problemas 8.43 y 8.44.

2. **PRUEBA DE HIPÓTESIS PARA VALORES PREDICHOS**. Sea y_0 el valor predicho de y correspondiente a $x = x_0$, como se estimó a partir de la ecuación de regresión muestral, es decir, $y_0 = a + bx_0$. Sea y_p el valor predicho de y correspondiente a $x = x_0$ para la población. Entonces, el estadístico

$$t = \frac{(y_0 - y_p)\sqrt{n-2}}{s_{y.x}\sqrt{n + 1 + [n(x_0 - \bar{x})^2/s_x^2]}} \tag{47}$$

tiene distribución de Student con $n - 2$ grados de libertad. A partir de esto se pueden encontrar los límites de confianza para los valores predichos de la población. *Véase* el problema 8.45.

3. **PRUEBA DE HIPÓTESIS PARA VALORES MEDIOS PREDICHOS**. Sea y_0 el valor predicho de y correspondiente a $x = x_0$, como se estimó a partir de la ecuación de regresión muestral, es decir, $y_0 = a + bx_0$. Sea $\bar{y}_p$ el *valor medio* predicho de y correspondiente a $x = x_0$ para la población [es decir, $\bar{y}_p = E(Y \mid X = x_0)$]. Entonces, el estadístico

$$t = \frac{(y_0 - \bar{y}_p)\sqrt{n-2}}{s_{y.x}\sqrt{1 + [(x_0 - \bar{x})^2/s_x^2]}} \tag{48}$$

tiene distribución de Student con $n - 2$ grados de libertad. A partir de esto se pueden encontrar los límites de confianza para los valores medios predichos de la población. *Véase* el problema 8.46.

TEORÍA MUESTRAL DE LA CORRELACIÓN

Con frecuencia debemos estimar el coeficiente de correlación de la población ρ a partir del coeficiente muestral de correlación r o probar la hipótesis que concierne ρ. Con este propósito debemos conocer la distribución muestral de r. En el caso $\rho = 0$, la distribución es simétrica y puede usarse el estadístico que tiene la distribución de Student. Para $\rho \neq 0$, la distribución está sesgada. En tal caso, una transformación debido a Fisher genera un estadístico cuya distribución es aproximadamente normal. Las siguientes pruebas resumen los procedimientos involucrados.

1. **PRUEBA DE HIPÓTESIS** $\rho = 0$. Aquí usamos el hecho de que el estadístico

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{49}$$

tiene la distribución de Student con $n - 2$ grados de libertad. *Véase* el problema 8.48.

2. PRUEBA DE HIPÓTESIS $\rho = \rho_0 \neq 0$. Aquí usamos el hecho de que el estadístico

$$Z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = 1.1513 \log_{10}\left(\frac{1+r}{1-r}\right) \tag{50}$$

tiene distribución aproximadamente normal con media y desviación estándar dadas por

$$\mu_z = \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right) = 1.1513 \log_{10}\left(\frac{1+\rho_0}{1-\rho_0}\right), \qquad \sigma_Z = \frac{1}{\sqrt{n-3}} \tag{51}$$

Estos hechos también se pueden usar para encontrar límites de confianza para los coeficientes de correlación. *Véanse* los problemas 8.49 y 8.50. La transformación (50) se llama la *transformación Z de Fisher*.

3. SIGNIFICANCIA DE LA DIFERENCIA ENTRE COEFICIENTES DE CORRELACIÓN. Para determinar si dos coeficientes de correlación r_1 y r_2 tomados de las muestras de tamaños n_1 y n_2, respectivamente, difieren significativamente entre sí, calculamos los Z_1 y Z_2 correspondientes a r_1 y r_2 usando (*50*). Luego, utilizamos el hecho de que el estadístico de la prueba

$$z = \frac{Z_1 - Z_2 - \mu_{Z_1 - Z_2}}{\sigma_{Z_1 - Z_2}} \tag{52}$$

donde $$\mu_{Z_1 - Z_2} = \mu_{Z_1} - \mu_{Z_2}, \qquad \sigma_{Z_1 - Z_2} = \sqrt{\sigma_{Z_1}^2 + \sigma_{Z_2}^2} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}} \tag{53}$$

tiene distribución normal. *Véase* el problema 8.51.

CORRELACIÓN Y DEPENDENCIA

Siempre que dos variables aleatorias X y Y tengan un coeficiente de correlación ρ diferente de cero, sabemos (teorema 3-15, página 89) que son *dependientes* en el sentido probabilístico (es decir, su distribución conjunta no se factoriza en sus distribuciones marginales). Más aún, cuando $\rho \neq 0$, podemos usar una ecuación de la forma (39) para *predecir* el valor de Y a partir del valor de X.

Es importante comprender que "correlación" y "dependencia" en el sentido anterior no necesariamente implican una interdependencia causal directa de X y Y. Esto se ilustra en los siguientes ejemplos.

EJEMPLO 8.1. Sean X y Y variables aleatorias que representan estaturas y pesos de individuos. Aquí hay una interdependencia directa entre X y Y.

EJEMPLO 8.2. Si X representa los salarios anuales de los profesores mientras que Y representa la cantidad de crímenes, el coeficiente de correlación puede ser diferente de cero y podremos estar en capacidad de encontrar una ecuación de regresión que prediga una variable a partir de la otra. Pero difícilmente vamos a querer afirmar que existe una interdependencia directa entre X y Y.

Problemas resueltos

LA RECTA DE MÍNIMOS CUADRADOS

8.1. Una línea recta pasa por los puntos (x_1, y_1) y (x_2, y_2). Demuestre que la ecuación de la línea es

$$y - y_1 = \left(\frac{y_2 - y_1}{x_2 - x_1}\right)(x - x_1)$$

La ecuación de la recta es $y = a + bx$. Entonces, ya que (x_1, y_1) y (x_2, y_2) son puntos sobre la recta, tenemos

$$y_1 = a + bx_1, \qquad y_2 = a + bx_2$$

Por consiguiente,

(*1*) $$y - y_1 = (a + bx) - (a + bx_1) = b(x - x_1)$$

(*2*) $$y_2 - y_1 = (a + bx_2) - (a + bx_1) = b(x_2 - x_1)$$

Obteniendo $b = (y_2 - y_1)/(x_2 - x_1)$ de (*2*) y sustituyendo en (*1*), se deduce el resultado requerido.

En la figura 8-5 se muestra la gráfica de la recta PQ. La constante $b = (y_2 - y_1)/(x_2 - x_1)$ es la pendiente de la recta.

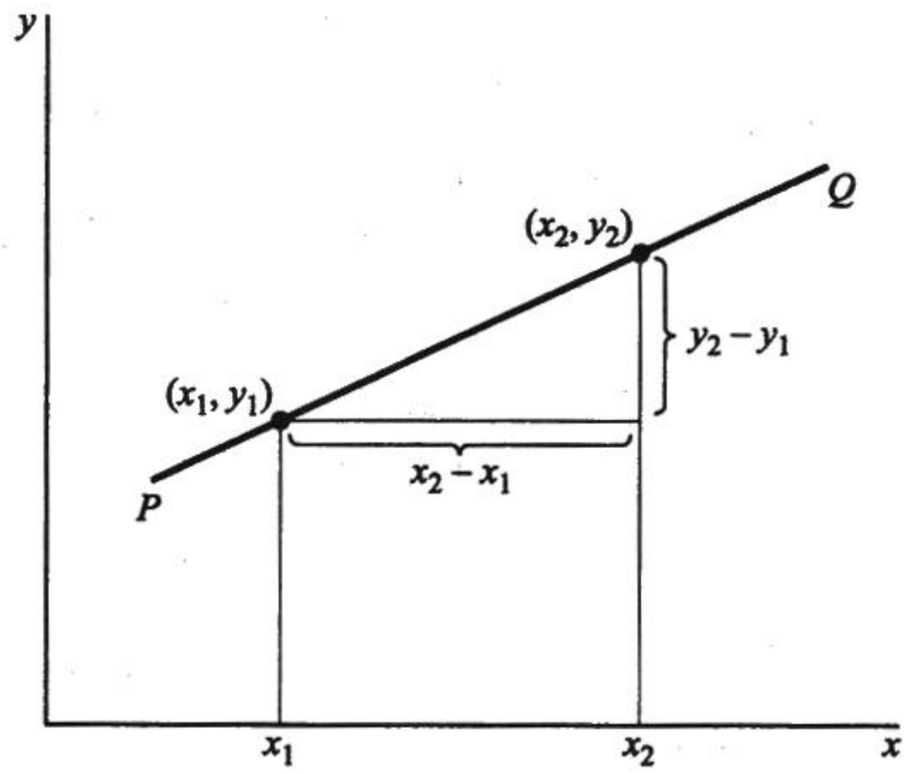


Fig. 8-5

8.2. (*a*) Construya una línea recta que se aproxime a los datos de la tabla 8-1. (*b*) Encuentre una ecuación para esta recta.

Tabla 8-1

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

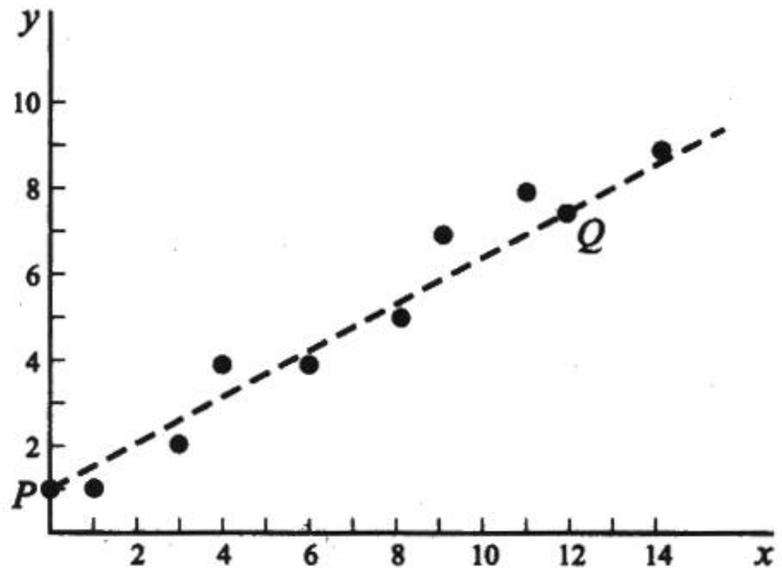


Fig. 8-6

(*a*) Dibuje los puntos (1, 1), (3, 2), (4, 4), (6, 4) (8, 5), (9, 7), (11, 8) y (14, 9) en un sistema de coordenadas rectangulares como se muestra en la figura 8-6.

Se dibuja una línea recta a mano alzada en la figura. Para un método que elimine la necesidad de juicio individual, *véase* el problema 8.4, el cual utiliza el método de mínimos cuadrados.

(*b*) Para obtener la ecuación de la recta dibujada en (*a*), escoja dos puntos cualesquiera de la recta, tales como P y Q. Las coordenadas de esos puntos, como se lee de la gráfica, son aproximadamente (0, 1) y (12, 7.5). Entonces, a partir del problema 8.1,

$$y - 1 = \frac{7.5 - 1}{12 - 0}(x - 0)$$

o $y - 1 = 0.542x$ o $y = 1 + 0.542x$.

8.3. Derive las ecuaciones normales (*4*), página 287, para la recta de mínimos cuadrados.

Refiérase a la figura 8-7. Los valores de y en la recta de mínimos cuadrados correspondientes a $x_1, x_2, \ldots, x_n$ son

$$a + bx_1, \quad a + bx_2, \quad \ldots, \quad a + bx_n$$

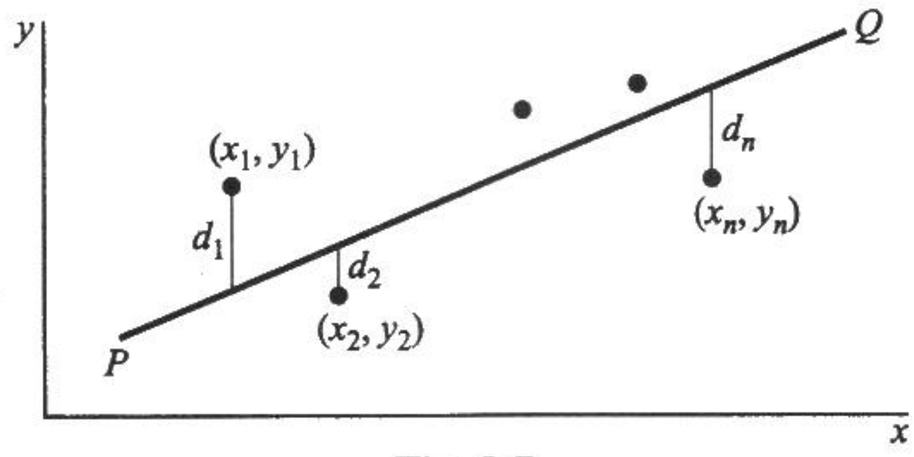


Fig. 8-7

Las desviaciones verticales correspondientes son

$$d_1 = a + bx_1 - y_1, \qquad d_2 = a + bx_2 - y_2, \qquad \ldots, \qquad d_n = a + bx_n - y_n$$

Entonces la suma de cuadrados de las desviaciones es

$$d_1^2 + d_2^2 + \cdots + d_n^2 = (a + bx_1 - y_1)^2 + (a + bx_2 - y_2)^2 + \cdots + (a + bx_n - y_n)^2$$

o

$$\sum d^2 = \sum (a + bx - y)^2$$

Ésta es una función de a y b, es decir, $F(a, b) = \sum(a + bx - y)^2$. Una condición necesaria para que esto sea un mínimo (o un máximo) es que $\partial F/\partial a = 0$, $\partial F/\partial b = 0$. Ya que

$$\frac{\partial F}{\partial a} = \sum \frac{\partial}{\partial a}(a + bx - y)^2 = \sum 2(a + bx - y)$$

$$\frac{\partial F}{\partial b} = \sum \frac{\partial}{\partial b}(a + bx - y)^2 = \sum 2x(a + bx - y)$$

obtenemos

$$\sum (a + bx - y) = 0 \qquad \sum x(a + bx - y) = 0$$

es decir,

$$\sum y = an + b \sum x \qquad \sum xy = a \sum x + b \sum x^2$$

como se requiere. Se puede demostrar que en realidad éstas resultan en un mínimo.

8.4. Ajuste una recta de mínimos cuadrados a los datos del problema 8.2 usando (*a*) x como variable independiente, (*b*) x como variable dependiente.

(*a*) La ecuación de la recta es $y = a + bx$. Las ecuaciones normales son

$$\sum y = an + b \sum x$$

$$\sum xy = a \sum x + b \sum x^2$$

El trabajo involucrado en calcular las sumas puede disponerse como en la tabla 8-2. Aunque la última columna no se necesita en esta parte del problema, se incluyó para usar la tabla en la parte (*b*).

Dado que hay 8 pares de valores x y y, $n = 8$ y la ecuación se convierte en

$$8a + 56b = 40$$

$$56a + 524b = 364$$

Resolviendo simultáneamente, $a = \frac{6}{11}$ o 0.545, $b = \frac{7}{11}$ o 0.636; y la recta de mínimos cuadrados requerida es $y = \frac{6}{11} + \frac{7}{11}x$ o $y = 0.545 + 0.636x$. Observe que ésta no es la recta que se obtuvo en el problema 8.2 usando el método a mano alzada.

Tabla 8-2

x	y	x^2	xy	y^2
1	1	1	1	1
3	2	9	6	4
4	4	16	16	16
6	4	36	24	16
8	5	64	40	25
9	7	81	63	49
11	8	121	88	64
14	9	196	126	81
$\sum x = 56$	$\sum y = 40$	$\sum x^2 = 524$	$\sum xy = 364$	$\sum y^2 = 256$

Otro método.

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n\sum x^2 - (\sum x)^2} = \frac{(40)(524) - (56)(364)}{(8)(524) - (56)^2} = \frac{6}{11} \quad \text{o} \quad 0.545$$

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{(8)(364) - (56)(40)}{(8)(524) - (56)^2} = \frac{7}{11} \quad \text{o} \quad 0.636$$

(*b*) Si x se considera como la variable dependiente y y como la variable independiente, la ecuación de la recta de mínimos cuadrados es $x = c + dy$ y las ecuaciones normales son

$$\sum x = cn + d\sum y$$

$$\sum xy = c\sum y + d\sum y^2$$

Entonces, usando la tabla 8-2, las ecuaciones normales se convierten en

$$8c + 40d = 56$$

$$40c + 256d = 364$$

a partir de lo cual $c = -\frac{1}{2}$ o -0.50, $d = \frac{3}{2}$ o 1.50.

Estos valores también pueden obtenerse de

$$c = \frac{(\sum x)(\sum y^2) - (\sum y)(\sum xy)}{n\sum y^2 - (\sum y)^2} = \frac{(56)(256) - (40)(364)}{(8)(256) - (40)^2} = -0.50$$

$$d = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum y^2 - (\sum y)^2} = \frac{(8)(364) - (56)(40)}{(8)(256) - (40)^2} = 1.50$$

Por consiguiente, la ecuación requerida de la recta de mínimos cuadrados es $x = -0.50 + 1.50y$.

Observe que resolviendo esta ecuación para y, obtenemos $y = 0.333 + 0.667x$, que no es la misma recta obtenida en la parte (*a*).

8.5. Dibuje las dos rectas obtenidas en el problema 8.4.

Las gráficas de las dos rectas, $y = 0.545 + 0.636x$ y $x = -0.500 + 1.50y$ se muestran en la figura 8-8. Observe que las dos rectas en este caso prácticamente coinciden, lo cual es un indicio de que los datos están muy bien descritos por una relación lineal.

La recta que se obtuvo en la parte (*a*) se denomina *recta de regresión de y en x* y se usa para estimar y para valores dados de x. La recta que se obtuvo en la parte (*b*) se llama *recta de regresión de x en y* y se usa para estimar x para valores dados de y.

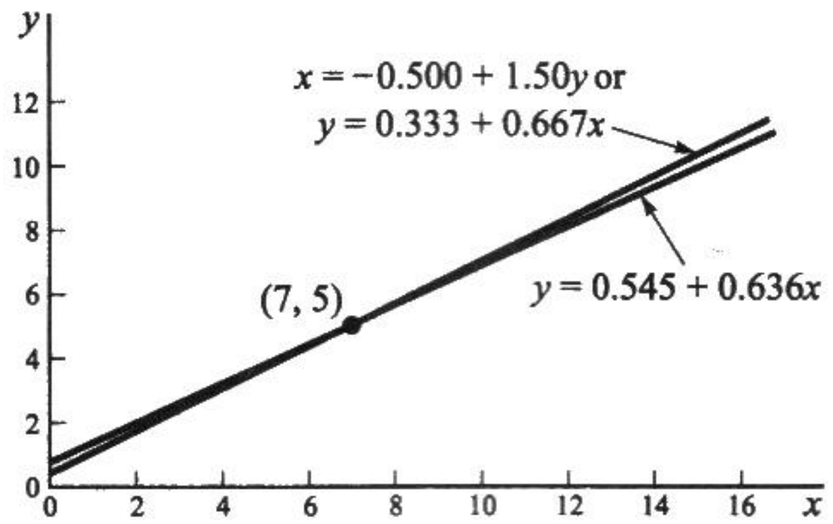


Fig. 8-8

8.6. (*a*) Demuestre que las dos rectas de mínimos cuadrados obtenidas en el problema 8.4 se intersecan en el punto $(\bar{x}, \bar{y})$. (*b*) Calcule el valor de y cuando $x = 12$. (*c*) Calcule el valor de x cuando $y = 3$.

$$\bar{x} = \frac{\sum x}{n} = \frac{56}{8} = 7, \qquad \bar{y} = \frac{\sum y}{n} = \frac{40}{8} = 5$$

Entonces el punto $(\bar{x}, \bar{y})$, llamado el *centroide*, es (7, 5).

(*a*) El punto (7, 5) está ubicado en la recta $y = 0.545 + 0.636x$ o, más exactamente, $y = \frac{6}{11} + \frac{7}{11}x$, ya que $5 = \frac{6}{11} + \frac{7}{11}(7)$.

El punto (7, 5) está ubicado en la recta $x = -\frac{1}{2} + \frac{3}{2}y$, ya que $7 = -\frac{1}{2} + \frac{3}{2}(5)$.

Otro método.

Las ecuaciones de las dos rectas son $y = \frac{6}{11} + \frac{7}{11}x$ y $x = -\frac{1}{2} + \frac{3}{2}y$. Resolviendo simultáneamente, encontramos que $x = 7$ y $y = 5$. Por lo tanto, las rectas se intersecan en el punto (7, 5).

(*b*) Remplazando $x = 12$ en la recta de regresión de y en x, $y = 0.545 + 0.636(12) = 8.2$.

(*c*) Remplazando $y = 3$ en la recta de regresión de x en y, $x = -0.50 + 1.50(3) = 4.0$.

8.7. Demuestre que una recta de mínimos cuadrados siempre atraviesa el punto $(\bar{x}, \bar{y})$.

Caso 1.

x es la variable independiente.

Una ecuación de la recta de mínimos cuadrados es (*1*) $y = a + bx$

La ecuación normal para la recta de mínimos cuadrados es (*2*) $\sum y = an + b \sum x$

Dividiendo ambos lados de (*2*) por n da (*3*) $\bar{y} = a + b\bar{x}$

Restando (*3*) de (*1*), la recta de mínimos cuadrados se puede escribir

(*4*) $$y - \bar{y} = b(x - \bar{x})$$

lo cual demuestra que la recta atraviesa el punto $(\bar{x}, \bar{y})$.

Caso 2.

y es la variable dependiente.

Procediendo tal como en el caso 1 con x y y intercambiadas y las constantes a, b remplazadas por c, d, respectivamente, encontramos que la recta de mínimos cuadrados se puede escribir

(*5*) $$x - \bar{x} = d(y - \bar{y})$$

lo cual indica que la recta atraviesa el punto $(\bar{x}, \bar{y})$.

Observe que, en general, las rectas (*4*) y (*5*) no coinciden, pero intersecan el punto $(\bar{x}, \bar{y})$.

8.8. Demuestre que la recta de regresión de mínimos cuadrados de y en x se puede escribir de la forma de (*8*), página 287.

Tenemos (*4*) del problema 8.7, $y - \bar{y} = b(x - \bar{x})$. A partir de la segunda ecuación en (*5*), página 287, tenemos

(*1*) $$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

Ahora
$$\begin{aligned}\sum (x-\bar{x})^2 &= \sum (x^2 - 2\bar{x}x + \bar{x}^2)\\ &= \sum x^2 - 2\bar{x}\sum x + \sum \bar{x}^2\\ &= \sum x^2 - 2n\bar{x}^2 + n\bar{x}^2\\ &= \sum x^2 - n\bar{x}^2\\ &= \sum x^2 - \frac{1}{n}(\sum x)^2\\ &= \frac{1}{n}[n\sum x^2 - (\sum x)^2]\end{aligned}$$

También
$$\begin{aligned}\sum (x-\bar{x})(y-\bar{y}) &= \sum (xy - \bar{x}y - \bar{y}x + \bar{x}\bar{y})\\ &= \sum xy - \bar{x}\sum y - \bar{y}\sum x + \sum \bar{x}\bar{y}\\ &= \sum xy - n\bar{x}\bar{y} - n\bar{y}\bar{x} + n\bar{x}\bar{y}\\ &= \sum xy - n\bar{x}\bar{y}\\ &= \sum xy - \frac{(\sum x)(\sum y)}{n}\\ &= \frac{1}{n}[n\sum xy - (\sum x)(\sum y)]\end{aligned}$$

Por consiguiente, (*1*) se convierte en

$$b = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$$

A partir de lo cual se obtiene el resultado (*8*). La demostración de (*12*), página 288, se deduce al intercambiar x y y.

8.9. Sean $x = x' + h$, $y = y' + k$, donde h y k son constantes cualesquiera. Demuestre que

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{n\sum x'y' - (\sum x')(\sum y')}{n\sum x'^2 - (\sum x')^2}$$

A partir del problema 8.8 tenemos

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$$

Ahora si $x = x' + h$, $y = y' + k$, tenemos

$$\bar{x} = \overline{x'} + h, \qquad \bar{y} = \overline{y'} + k$$

Por lo tanto
$$\begin{aligned}\frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} &= \frac{\sum (x' - \overline{x'})(y' - \overline{y'})}{\sum (x' - \overline{x'})^2}\\ &= \frac{n\sum x'y' - (\sum x')(\sum y')}{n\sum x'^2 - (\sum x')^2}\end{aligned}$$

El resultado es útil para desarrollar un método abreviado con el fin de obtener rectas de mínimos cuadrados sustrayendo constantes apropiadas de los valores dados de x y y (*véase* el problema 8.12).

8.10. Si, en particular, $h = \bar{x}$, $k = \bar{y}$ en el problema 8.9, demuestre que

$$b = \frac{\sum x'y'}{\sum x'^2}$$

Esto se deduce inmediatamente del problema 8.9 ya que

$$\sum x' = \sum (x - \bar{x}) = \sum x - n\bar{x} = 0$$

e igualmente $\sum y' = 0$.

8.11. La tabla 8-3 muestra las estaturas respectivas x y y de una muestra de 12 padres y sus hijos mayores. (a) Elabore un diagrama de dispersión. (b) Encuentre la recta de regresión de mínimos cuadrados de y en x. (c) Encuentre la recta de regresión de mínimos cuadrados de x en y.

Tabla 8-3

Estatura x del padre (pulgadas)	65	63	67	64	68	62	70	66	68	67	69	71
Estatura y del hijo (pulgadas)	68	66	68	65	69	66	68	65	71	67	68	70

(a) El diagrama de dispersión se obtiene dibujando los puntos (x, y) en un sistema rectangular de coordenadas, como se muestra en la figura 8-9.

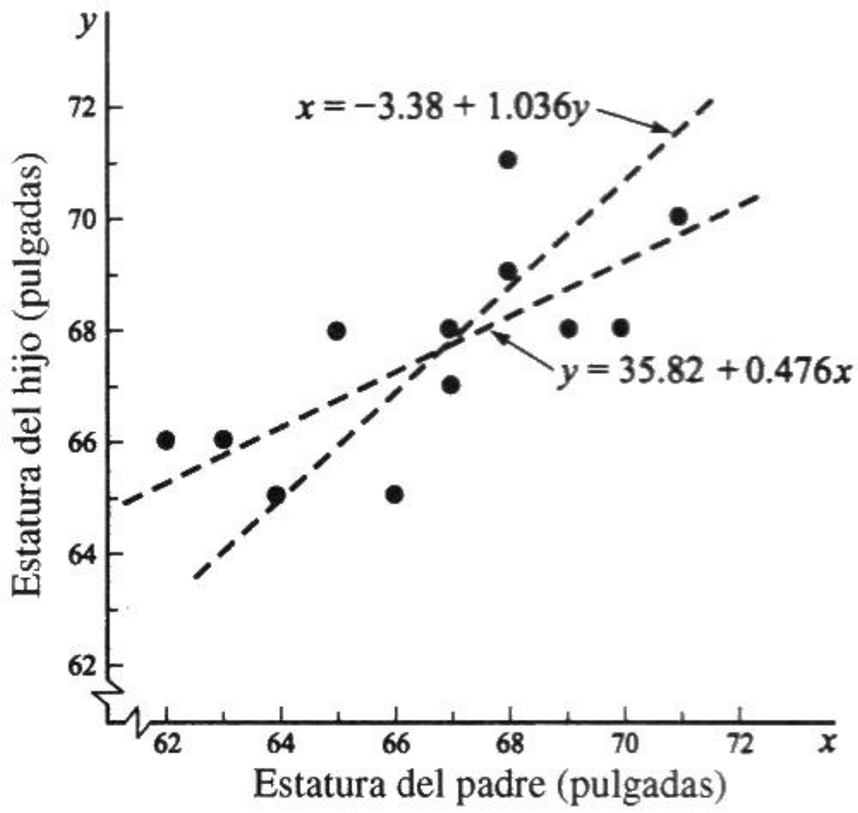


Fig. 8-9

(b) La recta de regresión de y en x está dada por $y = a + bx$, donde a y b se obtienen resolviendo las ecuaciones normales

$$\sum y = an + b \sum x$$

$$\sum xy = a \sum x + b \sum x^2$$

Las sumas se muestran en la tabla 8-4, y así las ecuaciones se convierten en

$$12a + 800b = 811$$

$$800a + 53{,}418b = 54{,}107$$

a partir de lo cual encontramos $a = 35.82$ y $b = 0.476$, de manera que $y = 35.82 + 0.476x$. La gráfica de esta ecuación se muestra en la figura 8-9.

Otro método.

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n \sum x^2 - (\sum x)^2} = 35.82, \qquad b = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2} = 0.476$$

Tabla 8-4

x	y	x^2	xy	y^2
65	68	4225	4420	4624
63	66	3969	4158	4356
67	68	4489	4556	4624
64	65	4096	4160	4225
68	69	4624	4692	4761
62	66	3844	4092	4356
70	68	4900	4760	4624
66	65	4356	4290	4225
68	71	4624	4828	5041
67	67	4489	4489	4489
69	68	4761	4692	4624
71	70	5041	4970	4900
$\sum x = 800$	$\sum y = 811$	$\sum x^2 = 53{,}418$	$\sum = 54{,}107$	$\sum y^2 = 54{,}849$

(*c*) La recta de regresión de x en y está dada por $x = c + dy$, donde c y d se obtienen resolviendo las ecuaciones normales

$$\sum x = cn + d \sum y$$

$$\sum xy = c \sum y + d \sum y^2$$

Usando las sumas de la tabla 8-4, éstas se convierten en

$$12c + 811d = 800$$

$$811c + 54{,}849d = 54{,}107$$

A partir de lo cual encontramos $c = -3.38$ y $d = 1.036$, de manera que $x = -3.38 + 1.036y$. La gráfica de esta ecuación aparece en la figura 8-9.

Otro método.

$$c = \frac{(\sum x)(\sum y^2) - (\sum y)(\sum xy)}{n \sum y^2 - (\sum y)^2} = -3.38, \qquad d = \frac{n \sum xy - (\sum y)(\sum x)}{n \sum y^2 - (\sum y)^2} = 1.036$$

8.12. Desarrolle el problema 8.11 usando el método del problema 8.9.

Reste un valor apropiado, digamos 68, a x y y (los números sustraídos a x y y pueden ser diferentes). Esto nos lleva a la tabla 8-5.

En la tabla encontramos

$$b = \frac{n \sum x'y' - (\sum x')(\sum y')}{n \sum x'^2 - (\sum x')^2} = \frac{(12)(47) - (-16)(-5)}{(12)(106) - (16)^2} = 0.476$$

También, ya que $x' = x - 68$, $y' = y - 68$, tenemos $\overline{x'} = \bar{x} - 68$, $\overline{y'} = \bar{y} - 68$. Así

$$\bar{x} = \overline{x'} + 68 = -\frac{16}{12} + 68 = 66.67, \qquad \bar{y} = \overline{y'} + 68 = -\frac{5}{12} + 68 = 67.58$$

La ecuación de regresión requerida de y en x es $y - \bar{y} = b(x - \bar{x})$, es decir,

$$y - 67.58 = 0.476(x - 66.07) \quad \text{o} \quad y = 35.85 + 0.476x$$

en concordancia con el problema 8.11, aparte de los errores al redondear las cifras. De manera similar, obtenemos la ecuación de regresión de x en y.

Tabla 8-5

x'	y'	x'^2	$x'y'$	y'^2
−3	0	9	0	0
−5	−2	25	10	4
−1	0	1	0	0
−4	−3	16	12	9
0	1	0	0	1
−6	−2	36	12	4
2	0	4	0	0
−2	−3	4	6	9
0	3	0	0	9
−1	−1	1	1	1
1	0	1	0	0
3	2	9	6	4
$\sum x' = -16$	$\sum y' = -5$	$\sum x'^2 = 106$	$\sum x'y' = 47$	$\sum y'^2 = 41$

ECUACIONES NO LINEALES REDUCIBLES A LA FORMA LINEAL

8.13. La tabla 8-6 da los valores experimentales de la presión P de una masa dada de gas correspondiente a varios valores del volumen V. Según los principios de la termodinámica, una relación que tiene la forma $PV^{\gamma} = C$, donde γ y C son constantes, debe existir entre las variables. (*a*) Encuentre los valores de γ y C. (*b*) Escriba una ecuación que conecte P y V. (*c*) Calcule P cuando $V = 100.0$ pulg^3.

Tabla 8-6

Volumen V (pulg^3)	54.3	61.8	72.4	88.7	118.6	194.0
Presión P (lb/pulg^2)	61.2	49.5	37.6	28.4	19.2	10.1

Dado que $PV^{\gamma} = C$, tenemos, tomando los logaritmos base 10,

$$\log P + \gamma \log V = \log C \qquad \text{o} \qquad \log P = \log C - \gamma \log V$$

Remplazando $V = x$ y $\log P = y$, la última ecuación se puede escribir

$$y = a + bx \tag{1}$$

donde $a = \log C$ y $b = -\gamma$

La tabla 8-7 da los valores de x y y correspondientes a los valores de V y P en la tabla 8-6, y también indica los cálculos involucrados al calcular la recta de mínimos cuadrados (*1*).

Tabla 8-7

$x = \log V$	$y = \log P$	x^2	xy
1.7348	1.7868	3.0095	3.0997
1.7910	1.6946	3.2077	3.0350
1.8597	1.5752	3.4585	2.9294
1.9479	1.4533	3.7943	2.8309
2.0741	1.2833	4.3019	2.6617
2.2878	1.0043	5.2340	2.2976
$\sum x = 11.6953$	$\sum y = 8.7975$	$\sum x^2 = 23.0059$	$\sum xy = 16.8543$

Las ecuaciones normales correspondientes a la recta de mínimos cuadrados (*1*) son

$$\sum y = an + b \sum x \qquad \sum xy = a \sum x + b \sum x^2$$

a partir de lo cual

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n \sum x^2 - (\sum x)^2} = 4.20, \qquad b = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2} = -1.40$$

Entonces $y = 4.20 - 1.40x$.

(*a*) Dado que $a = 4.20 = \log C$ y $b = -1.40 = -\gamma$, $C = 1.60 \times 10^4$ y $\gamma = 1.40$.

(*b*) $PV^{1.40} = 16{,}000$

(*c*) Cuando $V = 100$, $x = \log V = 2$ y $y = \log P = 4.20 - 1.40(2) = 1.40$. Entonces $P = \text{antilog } 1.40 = 25.1$ lb/pulg^2.

8.14. Resuelva el problema 8.13 dibujando los datos en papel log-log.

Para cada par de valores de la presión P y el volumen V en la tabla 8-6, obtenemos un punto que se dibuja sobre papel logarítmico especial: *papel log-log*, como se muestra en la figura 8-10.

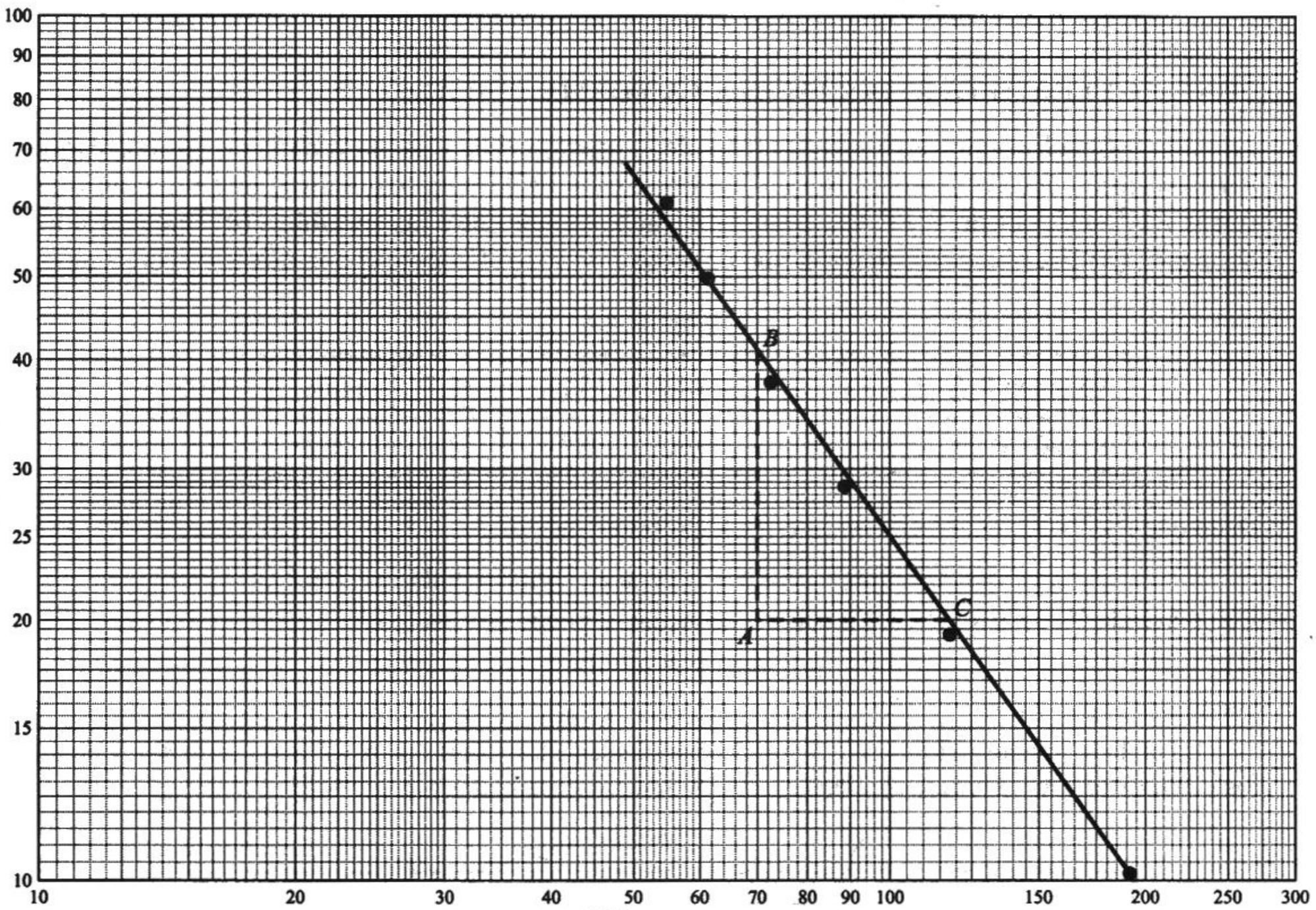


Fig. 8-10

También se indica una línea dibujada a mano alzada que aproxima estos puntos. La gráfica resultante muestra que hay una relación lineal entre $\log P$ y $\log V$, que se puede representar por la ecuación

$$\log P = a + b \log V \quad \text{o} \quad y = a + bx$$

La pendiente b, que en este caso es negativa, se da numéricamente por la relación de la longitud de AB a la longitud AC. En este caso la medida da $b = -1.4$.

Para obtener a, se necesita un punto de la recta. Por ejemplo, cuando $V = 100$, $P = 25$, de la gráfica. Entonces,

$$a = \log P - b \log V = \log 25 + 1.4 \log 100 = 1.4 + (1.4)(2) = 4.2$$

de manera que

$$\log P + 1.4 \log V = 4.2, \qquad \log PV^{1.4} = 4.2 \qquad \text{y} \qquad PV^{1.4} = 16{,}000$$

LA PARÁBOLA DE MÍNIMOS CUADRADOS

8.15. Derive las ecuaciones normales (*19*), página 289, para la parábola de mínimos cuadrados.

$$y = a + bx + cx^2$$

Sean los puntos muestrales (x_1, y_1), (x_2, y_2), . . ., (x_n, y_n). Entonces los valores de y en la parábola de mínimos cuadrados correspondientes a $x_1, x_2, \ldots, x_n$ son

$$a + bx_1 + cx_1^2, \qquad a + bx_2 + cx_2^2, \qquad \ldots, \qquad a + bx_n + cx_n^2$$

Por consiguiente, las desviaciones de $y_1, y_2, \ldots, y_n$ están dadas por

$$d_1 = a + bx_1 + cx_1^2 - y_1, \qquad d_2 = a + bx_2 + cx_2^2 - y_2, \qquad \ldots, \qquad d_n = a + bx_n + cx_n^2 - y_n$$

y la suma de los cuadrados de las desviaciones está dada por

$$\sum d^2 = \sum (a + bx + cx^2 - y)^2$$

Ésta es una función de a, b y c, es decir,

$$F(a, b, c) = \sum (a + bx + cx^2 - y)^2$$

Para minimizar la función, debemos tener

$$\frac{\partial F}{\partial a} = 0, \qquad \frac{\partial F}{\partial b} = 0, \qquad \frac{\partial F}{\partial c} = 0$$

Ahora

$$\frac{\partial F}{\partial a} = \sum \frac{\partial}{\partial a} (a + bx + cx^2 - y)^2 = \sum 2(a + bx + cx^2 - y)$$

$$\frac{\partial F}{\partial b} = \sum \frac{\partial}{\partial b} (a + bx + cx^2 - y)^2 = \sum 2x(a + bx + cx^2 - y)$$

$$\frac{\partial F}{\partial c} = \sum \frac{\partial}{\partial c} (a + bx + cx^2 - y)^2 = \sum 2x^2(a + bx + cx^2 - y)$$

Simplificando cada una de estas sumatorias e igualándolas a cero, se obtiene la ecuación (*19*), página 289.

8.16. Ajuste la parábola de mínimos cuadrados cuya forma es $y = a + bx + cx^2$ a los datos de la tabla 8-8.

Tabla 8-8

x	1.2	1.8	3.1	4.9	5.7	7.1	8.6	9.8
y	4.5	5.9	7.0	7.8	7.2	6.8	4.5	2.7

Entonces las ecuaciones normales son

$$\sum y = an + b \sum x + c \sum x^2$$

$$\sum xy = a \sum x + b \sum x^2 + c \sum x^3 \qquad (1)$$

$$\sum x^2 y = a \sum x^2 + b \sum x^3 + c \sum x^4$$

El trabajo involucrado en calcular las sumas puede disponerse como en la tabla 8-9.

Tabla 8-9

x	y	x^2	x^3	x^4	xy	x^2y
1.2	4.5	1.44	1.73	2.08	5.40	6.48
1.8	5.9	3.24	5.83	10.49	10.62	19.12
3.1	7.0	9.61	29.79	92.35	21.70	67.27
4.9	7.8	24.01	117.65	576.48	38.22	187.28
5.7	7.2	32.49	185.19	1055.58	41.04	233.93
7.1	6.8	50.41	357.91	2541.16	48.28	342.79
8.6	4.5	73.96	636.06	5470.12	38.70	332.82
9.8	2.7	96.04	941.19	9223.66	26.46	259.31
$\sum x =$ 42.2	$\sum y =$ 46.4	$\sum x^2 =$ 291.20	$\sum x^3 =$ 2275.35	$\sum x^4 =$ 18,971.92	$\sum xy =$ 230.42	$\sum x^2y =$ 1449.00

Entonces, dado que $n = 8$, las ecuaciones normales (*1*) se convierten en

$$\begin{aligned} 8a + 42.2b + 291.20c &= 46.4 \\ 42.2a + 291.20b + 2275.35c &= 230.42 \\ 291.20a + 2275.35b + 18971.92c &= 1449.00 \end{aligned} \tag{2}$$

Resolviendo $a = 2.588$, $b = 2.065$, $c = -0.2110$; así la parábola de mínimos cuadrados requerida tiene la ecuación

$$y = 2.588 + 2.065x - 0.2110x^2$$

8.17. Use la parábola de mínimos cuadrados de problema 8.16 para estimar los valores de y a partir de los valores dados de x.

Para $x = 1.2$, $y_{est} = 2.588 + 2.065(1.2) - 0.2110(1.2)^2 = 4.762$. De igual manera se obtienen otros valores estimados. Los resultados se muestran en la tabla 8-10 junto con los valores observados de y.

Tabla 8-10

y_{est}	4.762	5.621	6.962	7.640	7.503	6.613	4.741	2.561
y	4.5	5.9	7.0	7.8	7.2	6.8	4.5	2.7

REGRESIÓN MÚLTIPLE

8.18. Se va a estimar la variable z a partir de las variables x y y mediante una ecuación de regresión que tiene la forma $z = a + b + bx + cy$. Demuestre que la ecuación de regresión de mínimos cuadrados se obtiene determinando a, b y c de manera que satisfagan (*21*), página 289.

Sean los puntos muestrales $(x_1, y_1, z_1), \ldots, (x_n, y_n, z_n)$. Entonces los valores de z en el plano de regresión de mínimos cuadrados correspondientes a $(x_1, y_1), \ldots, (x_n, y_n)$ son, respectivamente

$$a + bx_1 + cy_1, \quad \ldots, \quad a + bx_n + cy_n$$

Por consiguiente, las desviaciones de $z_1, \ldots, z_n$ están dadas por

$$d_1 = a + bx_1 + cy_1 - z_1, \quad \ldots, \quad d_n = a + bx_n + cy_n - z_n$$

y la suma de los cuadrados de las desviaciones está dada por

$$\sum d^2 = \sum (a + bx + cy - z)^2$$

Considerando esto como una función de a, b, c y remplazando las derivadas parciales con respecto a a, b y c igual a cero, se obtienen las ecuaciones normales requeridas (*21*) en la página 289.

8.19. La tabla 8-11 muestra los pesos z aproximados a la libra más cercana, las estaturas x aproximadas a la pulgada más cercana y las edades y aproximadas al año más cercano de 12 niños. (*a*) Encuentre la ecuación de regresión de mínimos cuadrados de z en x y y. (*b*) Determine los valores estimados de z a partir de valores dados de x y y. (*c*) Estime el peso de un niño que tiene 9 años y su estatura es de 54 pulgadas.

Tabla 8-11

Peso (z)	64	71	53	67	55	58	77	57	56	51	76	68
Estatura (x)	57	59	49	62	51	50	55	48	52	42	61	57
Edad (y)	8	10	6	11	8	7	10	9	10	6	12	9

(*a*) La ecuación de regresión lineal de z en x y y se puede escribir

$$z = a + bx + cy$$

Las ecuaciones normales (*21*), página 289, están dadas por

$$\sum z = na + b\sum x + c\sum y$$

$$\sum xz = a\sum x + b\sum x^2 + c\sum xy \qquad (1)$$

$$\sum yz = a\sum y + b\sum xy + c\sum y^2$$

El trabajo involucrado en calcular las sumas puede presentarse como en la tabla 8-12.

Tabla 8-12

z	x	y	z^2	x^2	y^2	xz	yx	xy
64	57	8	4096	3249	64	3648	512	456
71	59	10	5041	3481	100	4189	710	590
53	49	6	2809	2401	36	2597	318	294
67	62	11	4489	3844	121	4154	737	682
55	51	8	3025	2601	64	2805	440	408
58	50	7	3364	2500	49	2900	406	350
77	55	10	5929	3025	100	4235	770	550
57	48	9	3249	2304	81	2736	513	432
56	52	10	3136	2704	100	2912	560	520
51	42	6	2601	1764	36	2142	306	252
76	61	12	5776	3721	144	4636	912	732
68	57	9	4624	3249	81	3876	612	513
$\sum z =$ 753	$\sum x =$ 643	$\sum y =$ 106	$\sum z^2 =$ 48,139	$\sum x^2 =$ 34,843	$\sum y^2 =$ 976	$\sum xz =$ 40,830	$\sum yz =$ 6796	$\sum xy =$ 5779

Usando esta tabla, las ecuaciones normales (*1*) se convierten en

$$12a + 643b + 106c = 753$$

$$643a + 34{,}843b + 5779c = 40{,}830 \qquad (2)$$

$$106a + 5779b + 976c = 6796$$

Resolviendo, $a = 3.6512$, $b = 0.8546$, $c = 1.5063$ y la ecuación de regresión requerida es

$$z = 3.65 + 0.855x + 1.506y \tag{3}$$

(*b*) Usando la ecuación de regresión (*3*), obtenemos los valores estimados de z, que se denotan por z_{est}, sustituyendo los valores correspondientes de x y y. Los resultados se dan en la tabla 8-13 junto con los valores muestrales de z.

Tabla 8-13

z_{est}	64.414	69.136	54.564	73.206	59.286	56.925	65.717	58.229	63.153	48.582	73.857	65.920
z	64	71	53	67	55	58	77	57	56	51	76	68

(*c*) Remplazando en (*3*) $x = 54$ y $y = 9$, el peso estimado es $z_{est} = 63.356$, o alrededor de 63 libras.

ERROR ESTÁNDAR DEL ESTIMADO

8.20. Si la recta de regresión de y en x está dada por $y = a + bx$, demuestre que el error estándar del estimado $s_{y \cdot x}$ está dado por

$$s_{y.x}^2 = \frac{\sum y^2 - a \sum y - b \sum xy}{n}$$

Los valores de y como estimado a partir de la recta de regresión están dados por $y_{est} = a + bx$. Entonces

$$s_{y.x}^2 = \frac{\sum (y - y_{est})^2}{n} = \frac{\sum (y - a - bx)^2}{n}$$

$$= \frac{\sum y(y - a - bx) - a \sum (y - a - bx) - b \sum x(y - a - bx)}{n}$$

Pero

$$\sum (y - a - bx) = \sum y - an - b \sum x = 0$$

$$\sum x(y - a - bx) = \sum xy - a \sum x - b \sum x^2 = 0$$

dado que las ecuaciones normales

$$\sum y = an + b \sum x \qquad \sum xy = a \sum x + b \sum x^2$$

Entonces

$$s_{y.x}^2 = \frac{\sum y(y - a - bx)}{n} = \frac{\sum y^2 - a \sum y - b \sum xy}{n}$$

Este resultado se puede extender a ecuaciones de regresión no lineal.

8.21. Demuestre que el resultado del problema 8.20 se puede escribir como

$$s_{y.x}^2 = \frac{\sum (y - \bar{y})^2 - b \sum (x - \bar{x})(y - \bar{y})}{n}$$

Método 1.

Sean $x = x' + \bar{x}$, $y = y' + \bar{y}$. Entonces, a partir del problema 8.20

$$\begin{aligned} ns_{y.x}^2 &= \sum y^2 - a\sum y - b\sum xy \\ &= \sum (y' + \bar{y})^2 - a\sum (y' + \bar{y}) - b\sum (x' + \bar{x})(y' + \bar{y}) \\ &= \sum (y'^2 + 2y'\bar{y} + \bar{y}^2) - a(\sum y' + n\bar{y}) - b\sum (x'y' + \bar{x}y' + x'\bar{y} + \bar{x}\bar{y}) \\ &= \sum y'^2 + 2\bar{y}\sum y' + n\bar{y}^2 - an\bar{y} - b\sum x'y' - b\bar{x}\sum y' - b\bar{y}\sum x' - bn\bar{x}\bar{y} \\ &= \sum y'^2 + n\bar{y}^2 - an\bar{y} - b\sum x'y' - bn\bar{x}\bar{y} \\ &= \sum y'^2 - b\sum x'y' + n\bar{y}(\bar{y} - a - b\bar{x}) \\ &= \sum y'^2 - b\sum x'y' \\ &= \sum (y - \bar{y})^2 - b\sum (x - \bar{x})(y - \bar{y}) \end{aligned}$$

donde usamos los resultados $\sum x' = 0$, $\sum y' = 0$ y $\bar{y} = a + b\bar{x}$ (lo cual se deduce al dividir ambos lados de la ecuación normal $\sum y = an + b\sum x$ por n). Esto demuestra el resultado requerido.

Método 2.

Sabemos que la recta de regresión se puede escribir como $y - \bar{y} = b(x - \bar{x})$, lo cual corresponde a empezar con $y = a + bx$ y luego remplazar a por cero, x por $x - \bar{x}$ y y por $y - \bar{y}$. Cuando estos remplazos se efectúan en el problema 8.20, se obtiene el resultado requerido.

8.22. Calcule el error estándar del estimado $S_{y.x}$ por los datos del problema 8.11.

Del problema 8.11 (b) la recta de regresión y en x es $y = 35.82 + 0.476x$. En la tabla 8-14 se muestran los valores reales de y (de la tabla 8-3) y los valores estimados de y, denotados por y_{est}, obtenidos de la recta de regresión. Por ejemplo, para el correspondiente $x = 65$, tenemos $y_{est} = 35.82 + 0.476(65) = 66.76$.

Tabla 8-14

x	65	63	67	64	68	62	70	66	68	67	69	71
y	68	66	68	65	69	66	68	65	71	67	68	70
y_{est}	66.76	65.81	67.71	66.28	68.19	65.33	69.14	67.24	68.19	67.71	68.66	69.62
$y - y_{est}$	1.24	0.19	0.29	−1.28	0.81	0.67	−1.14	−2.24	2.81	−0.71	−0.66	0.38

También se muestran los valores $y - y_{est}$, los cuales se necesitan para calcular $s_{y.x}$.

$$s_{y.x}^2 = \frac{\sum (y - y_{est})^2}{n} = \frac{(1.24)^2 + (0.19) + \cdots + (0.38)^2}{12} = 1.642$$

y $s_{y.x} = \sqrt{1.642} = 1.28$ pulgadas.

8.23. (a) Dibuje dos rectas paralelas a la recta de regresión del problema 8.11, cuya distancia vertical a ésta última sea $s_{y.x}$. (b) Determine el porcentaje de puntos de datos que hay entre las dos rectas.

(a) La recta de regresión $y = 35.82 + 0.476x$, como se obtuvo en el problema 8.11, aparece como una línea continua en la figura 8-11. Las dos rectas paralelas, cada una con distancia vertical $s_{y.x} = 1.28$ (*véase* problema 8.22) a la recta continua, aparecen punteadas en la figura 8-11.

(b) A partir de la figura se observa que de los 12 puntos de datos, 7 se encuentran entre las rectas, mientras que 3 caen sobre ellas. Un examen posterior usando el último renglón de la tabla 8-14 muestra que 2 de esos 3 puntos se encuentran entre las rectas. Entonces, el porcentaje requerido es $9/12 = 75\%$.

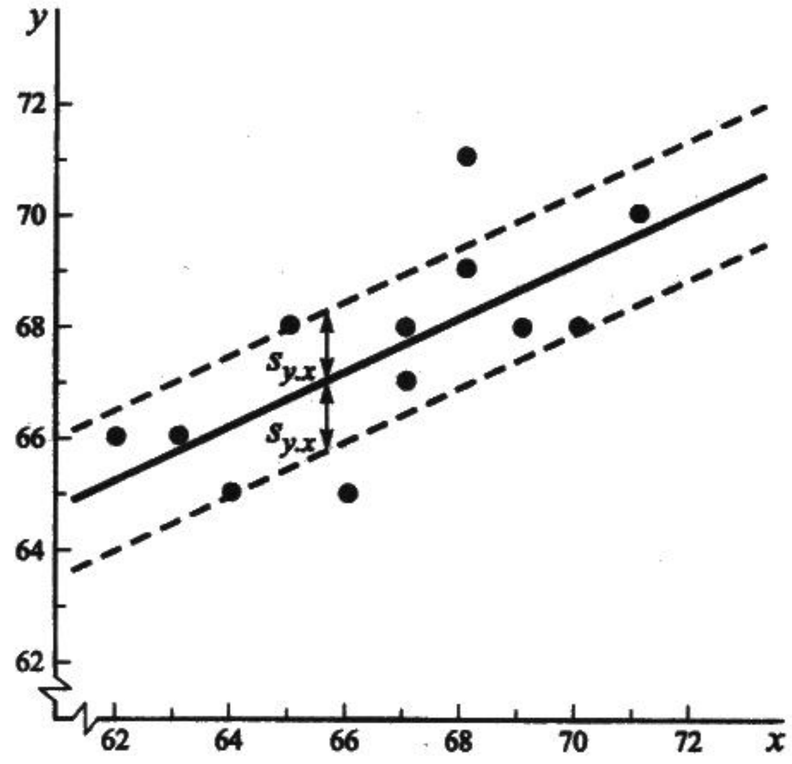


Fig. 8-11

Otro método.

Según el último renglón de la tabla 8-14, $y - y_{est}$ se encuentra entre -1.28 y 1.28 (es decir, $\pm s_{y.x}$) para 9 puntos (x, y). Entonces, el porcentaje requerido es $9/12 = 75\%$.

Si los puntos se distribuyen normalmente alrededor de la recta de regresión, la teoría predice que cerca del 68% de los puntos se ubica entre las rectas. Éste hubiera sido el caso si la muestra hubiese sido grande.

NOTA: Un mejor estimado del error estándar de la población a partir de la cual se tomaron las estaturas muestrales, está dado por $\hat{s}_{y.x} = \sqrt{n/(n-2)}s_{y.x} = \sqrt{12/10}(1.28) = 1.40$ pulgadas.

EL COEFICIENTE DE CORRELACIÓN LINEAL

8.24. Demuestre que $\sum (y - \bar{y})^2 = \sum (y - y_{est})^2 + \sum (y_{est} - \bar{y})^2$.

Elevando al cuadrado ambos lados de $y - \bar{y} = (y - y_{est}) + (y_{est} - \bar{y})$ y luego sumando, tenemos

$$\sum (y - \bar{y})^2 = \sum (y - y_{est})^2 + \sum (y_{est} - \bar{y})^2 + 2 \sum (y - y_{est})(y_{est} - \bar{y})$$

El resultado requerido se deduce inmediatamente si podemos demostrar que la última suma es cero. En el caso de la regresión lineal es así, ya que

$$\begin{aligned} \sum (y - y_{est})(y_{est} - \bar{y}) &= \sum (y - a - bx)(a + bx - \bar{y}) \\ &= a \sum (y - a - bx) + b \sum x(y - a - bx) - \bar{y} \sum (y - a - bx) \\ &= 0 \end{aligned}$$

debido a que las ecuaciones normales $\sum (y - a - bx) = 0$, $\sum x(y - a - bx) = 0$.

De manera similar se puede demostrar que el resultado es válido para una regresión no lineal usando una recta de mínimos cuadrados dada por $y_{est} = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$.

8.25. Calcule (a) la variación explicada, (b) la variación no explicada, (c) la variación total para los datos del problema 8.11.

Del problema 8.12 tenemos $\bar{y} = 67.58$ (o de la tabla 8-4, ya que $\bar{y} = 811/12 = 67.58$). Usando los valores y_{est} de la tabla 8-14 podemos construir la tabla 8-15.

Tabla 8-15

$y_{est} - \bar{y}$	−0.82	−1.77	0.13	−1.30	0.61	−2.25	1.56	−0.34	0.61	0.13	1.08	2.04

(a) Variación explicada $= \sum (y_{est} - \bar{y})^2 = (-0.82)^2 + \cdots + (2.04)^2 = 19.22$.

(b) Variación no explicada $= \sum (y - y_{est})^2 = ns_{y.x}^2 = 19.70$, del problema 8.22.

(*c*) Variación total $= \sum (y - \bar{y})^2 = 19.22 + 19.70 = 38.92$, del problema 8.24.

También se puede obtener el resultado en (*b*) y (*c*) haciendo directamente la suma de los cuadrados.

8.26. Encuentre (*a*) el coeficiente de determinación, (*b*) el coeficiente de correlación para los datos del problema 8.11. Use los resultados del problema 8.25.

(*a*) Coeficiente de determinación $= r^2 = \dfrac{\text{variación explicada}}{\text{variación total}} = \dfrac{19.22}{38.92} = 0.4938.$

(*b*) Coeficiente de correlación $= r = \pm\sqrt{0.4938} = \pm 0.7027.$

dado que la variable y_{est} aumenta a medida que x aumenta, la correlación es positiva y, por consiguiente, escribimos $r = 0.7027$ o 0.70 con dos cifras significativas.

8.27. A partir del resultado general (*30*), página 291, para el coeficiente de correlación, derive el resultado (*34*), página 291 (la *fórmula de producto-momento*), en el caso de regresión lineal.

La recta de regresión de mínimos cuadrados de y en x se puede escribir $y_{est} = a + bx$ o $y'_{est} = bx'$, donde $b = \sum x'y' / \sum x'^2$, $x' = x - \bar{x}$, y $y'_{est} = y_{est} - \bar{y}$. Entonces, usando $y' = y - \bar{y}$, tenemos

$$r^2 = \frac{\text{variación explicada}}{\text{variación total}} = \frac{\sum (y_{est} - \bar{y})^2}{\sum (y - \bar{y})^2} = \frac{\sum y'^2_{est}}{\sum y'^2}$$

$$= \frac{\sum b^2 x'^2}{\sum y'^2} = \frac{b^2 \sum x'^2}{\sum y'^2} = \left(\frac{\sum x'y'}{\sum x'^2}\right)^2 \left(\frac{\sum x'^2}{\sum y'^2}\right) = \frac{(\sum x'y')^2}{\sum x'^2 \sum y'^2}$$

y así

$$r = \pm \frac{\sum x'y'}{\sqrt{\sum x'^2 \sum y'^2}}$$

Sin embargo, ya que $\sum x'y'$ es positiva cuando y_{est} aumenta a medida que x aumenta, pero negativa cuando y_{est} disminuye a medida que x aumenta, la expresión para r tiene de manera automática el signo correcto asociado a ella. Por lo tanto, se obtiene el resultado requerido.

Tabla 8-16

x	y	$x' = x - \bar{x}$	$y' = y - \bar{y}$	x'^2	$x'y'$	y'^2
65	68	−1.7	0.4	2.89	−0.68	0.16
63	66	−3.7	−1.6	13.69	5.92	2.56
67	68	0.3	0.4	0.09	0.12	0.16
64	65	−2.7	−2.6	7.29	7.02	6.76
68	69	1.3	1.4	1.69	1.82	1.96
62	66	−4.7	−1.6	22.09	7.52	2.56
70	68	3.3	0.4	10.89	1.32	0.16
66	65	−0.07	−2.6	0.49	1.82	6.76
68	71	1.3	3.4	1.69	4.42	11.56
67	67	0.3	−0.6	0.09	−0.18	0.36
69	68	2.3	0.4	5.29	0.92	0.16
71	70	4.3	2.4	18.49	10.32	5.76
$\sum x = 800$ $\bar{x} = 800/12$ $= 66.7$	$\sum y = 811$ $\bar{y} = 811/12$ $= 67.6$			$\sum x'^2 =$ 84.68	$\sum x'y' =$ 40.34	$\sum y'^2 =$ 38.92

8.28. Usando la fórmula de producto-momento obtenga el coeficiente de correlación lineal para los datos del problema 8.11.

El trabajo que involucra el cálculo se puede organizar como lo muestra la tabla 8-16. Entonces,

$$r = \frac{\sum x'y'}{\sqrt{(\sum x'^2)(\sum y'^2)}} = \frac{40.34}{\sqrt{(84.68)(38.92)}} = 0.7027$$

lo cual concuerda con el problema 8.26(b).

8.29. Demuestre el resultado (*17*), página 288.

La recta de regresión de y en x es

$$y = a + bx, \quad \text{donde} \quad b = \frac{rs_y}{s_x}$$

De igual manera, la recta de regresión de x en y es

$$x = c + dy, \quad \text{donde} \quad d = \frac{rs_x}{s_y}$$

Entonces

$$bd = \left(\frac{rs_y}{s_x}\right)\left(\frac{rs_x}{s_y}\right) = r^2$$

8.30. Use el resultado del problema 8.29 para encontrar el coeficiente de correlación lineal para los datos del problema 8.11.

A partir del problema 8.11(b) y 8.11(c), respectivamente

$$b = \frac{484}{1016} = 0.476 \; ; \qquad d = \frac{484}{467} = 1.036$$

Entonces

$$r^2 = bd = \left(\frac{484}{1016}\right)\left(\frac{484}{467}\right) \; ; \quad \text{o} \quad r = 0.7027$$

lo que concuerda con los problemas 8.26(b) y 8.28.

8.31. Demuestre que el coeficiente de correlación lineal está dado por

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

En el problema 8.27 se demostró que

(*1*)
$$r = \frac{\sum x'y'}{\sqrt{(\sum x'^2)(\sum y'^2)}} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{[\sum (x-\bar{x})^2][\sum (y-\bar{y})^2]}}$$

pero

$$\begin{aligned}\sum (x-\bar{x})(y-\bar{y}) &= \sum (xy - \bar{x}y - x\bar{y} + \bar{x}\bar{y}) = \sum xy - \bar{x}\sum y - \bar{y}\sum x + n\bar{x}\bar{y} \\ &= \sum xy - n\bar{x}\bar{y} - n\bar{y}\bar{x} + n\bar{x}\bar{y} = \sum xy - n\bar{x}\bar{y} \\ &= \sum xy - \frac{(\sum x)(\sum y)}{n}\end{aligned}$$

ya que $\bar{x} = (\sum x)/n$ y $\bar{y} = (\sum y)/n$.

De igual manera,

$$\begin{aligned}\sum (x-\bar{x})^2 &= \sum (x^2 - 2x\bar{x} + \bar{x}^2) = \sum x^2 - 2\bar{x}\sum x + n\bar{x}^2 \\ &= \sum x^2 - \frac{2(\sum x)^2}{n} + \frac{(\sum x)^2}{n} = \sum x^2 - \frac{(\sum x)^2}{n}\end{aligned}$$

y $$\sum (y - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n}$$

Entonces, (*1*) se convierte en

$$r = \frac{\sum xy - (\sum x)(\sum y)/n}{\sqrt{[\sum x^2 - (\sum x)^2/n][\sum y^2 - (\sum y)^2/n]}} = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

8.32. Use la fórmula del problema 8.31 para obtener el coeficiente de correlación lineal para los datos del problema 8.11.

A partir de la tabla 8-4

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

$$= \frac{(12)(54{,}107) - (800)(811)}{\sqrt{[(12)(53{,}418) - (800)^2][(12)(54{,}849) - (811)^2]}} = 0.7027$$

tal como en los problemas 8.26(*b*), 8.28 y 8.30.

COEFICIENTE DE CORRELACIÓN GENERALIZADO

8.33. (*a*) Encuentre el coeficiente de correlación lineal entre las variables x y y del problema 8.16. (*b*) Encuentre el coeficiente de correlación no lineal entre esas variables, suponiendo la relación parabólica obtenida en el problema 8.16. (*c*) Explique la diferencia que hay entre los coeficientes de correlación obtenidos en (*a*) y (*b*). (*d*) ¿Qué porcentaje de la variación total permanece sin explicación debido a la suposición de la relación parabólica entre x y y?

(*a*) Usando los cálculos de la tabla 8-9 y el hecho de que $\sum y^2 = 290.52$, encontramos

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

$$= \frac{(8)(230.42) - (42.2)(46.4)}{\sqrt{[(8)(291.20) - (42.2)^2][(8)(290.52) - (46.4)^2]}} = -0.3743$$

(*b*) De la tabla 8-9, $\bar{y} = (\sum y)/n = (46.4)/8 = 5.80$. Entonces

$$\text{Variación total} = \sum (y - \bar{y})^2 = 21.40$$

De la tabla 8-10

$$\text{Variación explicada} = \sum (y_{\text{est}} - \bar{y})^2 = 21.02$$

Por consiguiente, $$r^2 = \frac{\text{variación explicada}}{\text{variación total}} = \frac{21.02}{21.40} = 0.9822 \qquad \text{y} \qquad r = 0.9911$$

(*c*) El hecho de que la parte (*a*) muestra un coeficiente de correlación de sólo -0.3743 indica que prácticamente no hay una *relación lineal* entre x y y. Sin embargo, hay una muy buena *relación no lineal* suministrada por la parábola del problema 8.16, como se indica por el hecho de que el coeficiente de correlación en (*b*) es muy cercano a 1.

(*d*) $$\frac{\text{Variación no explicada}}{\text{Variación total}} = 1 - r^2 = 1 - 0.9822 = 0.0178$$

Por lo tanto, 1.78% de la variación total permanece sin explicación. Esto se puede deber a fluctuaciones aleatorias o a una variable adicional que no se consideró.

8.34. Encuentre (*a*) S_y y (*b*) $S_{y\cdot x}$ para los datos del problema 8.16.

(*a*) A partir del problema 8.33(*b*), $\sum(y - \bar{y})^2 = 21.40$. Entonces la desviación estándar de y es

$$s_y = \sqrt{\frac{\sum (y - \bar{y})^2}{n}} = \sqrt{\frac{21.40}{8}} = 1.636 \text{ ó } 1.64$$

(*b*) **Primer método.**

Usando (*a*) y el problema 8.33(*b*), el error estándar del estimado de y en x es

$$s_{y\cdot x} = s_y\sqrt{1 - r^2} = 1.636\sqrt{1 - (0.9911)^2} = 0.218 \text{ ó } 0.22$$

Segundo método.

Usando el problema 8.33,

$$s_{y\cdot x} = \sqrt{\frac{\sum (y - y_{\text{est}})^2}{n}} = \sqrt{\frac{\text{variación no explicada}}{n}} = \sqrt{\frac{21.40 - 21.02}{8}} = 0.218 \text{ ó } 0.22$$

Tercer método.

Usando el problema 8.16 y el cálculo adicional $\sum y^2 = 290.52$, tenemos

$$s_{y\cdot x} = \sqrt{\frac{\sum y^2 - a\sum y - b\sum xy - c\sum x^2 y}{n}} = 0.218 \text{ ó } 0.22$$

8.35. Explique cómo podría usted determinar un coeficiente de correlación múltiple para las variables del problema 8.19.

Dado que z se determina a partir de x y y, estamos interesados en el coeficiente de correlación múltiple de z en x y en y. Para obtener esto vemos, a partir del problema 8.19, que

$$\text{Variación no explicada} = \sum (z - z_{\text{est}})^2$$
$$= (64 - 64.414)^2 + \cdots + (68 - 65.920)^2 = 258.88$$

$$\text{Variación total} = \sum (z - \bar{z})^2 = \sum z^2 - n\bar{z}^2$$
$$= 48{,}139 - 12(62.75)^2 = 888.25$$

$$\text{Variación explicada} = 888.25 - 258.88 = 629.37$$

Entonces

$$\text{El coeficiente de correlación múltiple de } z \text{ en } x \text{ y en } y$$
$$= \sqrt{\frac{\text{variación explicada}}{\text{variación total}}} = \sqrt{\frac{629.37}{888.25}} = 0.8418$$

Se debe mencionar que si fuésemos a considerar la regresión de x en y y en z, el coeficiente de correlación múltiple de x en y y en z podría, en general, ser diferente del valor anterior.

CORRELACIÓN DE RANGO

8.36. Derive la fórmula (*36*), página 292, de correlación de rango de Spearman.

Aquí estamos considerando n valores de x (a saber, pesos) y n valores correspondientes de y (a saber, estaturas). Sea x_j el de rango dado al j-ésimo valor de x y y_j el rango dado al j-ésimo valor de y. Los rangos son los enteros 1 hasta n. La media de x_j es entonces

$$\bar{x} = \frac{1 + 2 + \cdots + n}{n} = \frac{n(n+1)/2}{n} = \frac{n+1}{2}$$

mientras que la varianza

$$s_x^2 = \overline{x^2} - \bar{x}^2 = \frac{1^2 + 2^2 + \cdots + n^2}{n} - \left(\frac{n+1}{2}\right)^2$$

$$= \frac{n(n+1)(2n+1)/6}{n} - \left(\frac{n+1}{2}\right)^2$$

$$= \frac{n^2 - 1}{12}$$

usando los resultados 1 y 2 del apéndice A. De igual manera, la media $\bar{y}$ y la varianza s_y^2 son iguales a $(n + 1)/2$ y $(n^2 - 1)/12$, respectivamente.

Ahora, si $d_j = x_j - y_j$ son desviaciones entre los rangos, la varianza de las desviaciones, s_d^2, está dada en términos de s_x^2, s_y^2 y el coeficiente de correlación entre los rangos por

$$s_d^2 = s_x^2 + s_y^2 - 2r_{\text{rank}} s_x s_y$$

Entonces

$$r_{\text{rank}} = \frac{s_x^2 + s_y^2 - s_d^2}{2s_x s_y} \tag{1}$$

Dado que $\bar{d} = 0$, $s_d^2 = (\sum d^2)/n$ y (*1*) se convierte en

$$r_{\text{rank}} = \frac{(n^2 - 1)/12 + (n^2 - 1)/12 - (\sum d^2)/n}{(n^2 - 1)/6} = 1 - \frac{6\sum d^2}{n(n^2 - 1)} \tag{2}$$

8.37. La tabla 8-17 muestra cómo se clasificaron 10 estudiantes según su rendimiento en clase de biología en la parte teórica y en el laboratorio. Encuentre el coeficiente de correlación gradual.

Tabla 8-17

Laboratorio	8	3	9	2	7	10	4	6	1	5
Teoría	9	5	10	1	8	7	3	4	2	6

La diferencia de los rangos d en el laboratorio y la teoría para cada estudiante está dada en la tabla 8-18. Allí también se dan d^2 y $\sum d^2$.

Tabla 8-18

Diferencia de grados, d	−1	−2	−1	1	−1	3	1	2	−1	−1	
d^2	1	4	1	1	1	9	1	4	1	1	$\sum d^2 = 24$

Entonces

$$r_{\text{rank}} = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{6(24)}{10(10^2 - 1)} = 0.8545$$

lo que indica que hay una estrecha relación entre el desempeño en teoría y laboratorio.

8.38. Calcule el coeficiente de correlación de rango para los datos del problema 8.11 y compare su resultado con el coeficiente de correlación obtenido por otros métodos.

Colocadas en orden ascendente de magnitud, las estaturas de los padres son

(*1*) 62, 63, 64, 65, 66, 67, 67, 68, 68, 69, 70, 71

Ya que los lugares sexto y séptimo de este arreglo representan la misma estatura (67 pulgadas), asignamos un *rango medio* de 6.5 a ambos lugares. De igual manera, los lugares octavo y noveno tienen el rango 8.5. Por consiguiente, a las estaturas de los padres se les asignan los rangos.

(*2*) **1, 2, 3, 4, 5, 6.5, 6.5, 8.5, 8.5, 10, 11, 12**

Asimismo, las estaturas de los hijos dispuestas en orden ascendente de magnitud son

(*3*) **65, 65, 66, 66, 67, 68, 68, 68, 68, 69, 70, 71**

y dado que los lugares sexto, séptimo, octavo y noveno representan la misma estatura (pulgadas), asignamos el *rango medio* 7.5 [(6 + 7 + 8 + 9)/4] a estos lugares. Por consiguiente, a las estaturas de los hijos se les asigna los rangos

(*4*) **1.5, 1.5, 3.5, 3.5, 5, 7.5, 7.5, 7.5, 7.5, 10, 11, 12**

usando las correspondencias (*1*) y (*2*), (*3*) y (*4*), la tabla 8-3 se convierte en la tabla 8-19.

Tabla 8-19

Rango del padre	4	2	6.5	3	8.5	1	11	5	8.5	6.5	10	12
Rango del hijo	7.5	3.5	7.5	1.5	10	3.5	7.5	1.5	12	5	7.5	11

En la tabla 8-20 se muestran las diferencias en los rangos, d, y los cálculos de d^2 y $\sum d^2$.

Tabla 8-20

d	−3.5	−1.5	−1.0	1.5	−1.5	−2.5	3.5	3.5	−3.5	1.5	2.5	1.0	
d^2	12.25	2.25	1.00	2.25	2.25	6.25	12.25	12.25	12.25	2.25	6.25	1.00	$\sum d^2 = 72.50$

Entonces $$r_{\text{rank}} = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6(72.50)}{12(12^2-1)} = 0.7465$$

lo cual coincide bien con el valor $r = 0.7027$ obtenido en el problema 8.26(*b*).

INTERPRETACIÓN PROBABILÍSTICA DE LA REGRESIÓN Y LA CORRELACIÓN

8.39. Deduzca (*39*) a partir de (*37*).

Suponiendo que la ecuación de regresión es

$$y = E(Y \mid X = x) = \alpha + \beta x$$

Para la recta de regresión de mínimos cuadrados debemos considerar

$$\begin{aligned} E\{[Y-(\alpha+\beta X)]^2\} &= E\{[(Y-\mu_Y)-\beta(X-\mu_X)+(\mu_Y-\beta\mu_X-\alpha)]^2\} \\ &= E[(Y-\mu_Y)^2]+\beta^2 E[(X-\mu_X)^2]-2\beta E[(X-\mu_X)(Y-\mu_Y)]+(\mu_Y-\beta\mu_X-\alpha)^2 \\ &= \sigma_Y^2+\beta^2\sigma_X^2-2\beta\sigma_{XY}+(\mu_Y-\beta\mu_X-\alpha)^2 \end{aligned}$$

donde usamos $E(X - \mu_x) = 0$, $E(Y - \mu_y) = 0$

Denotando la última expresión como $F(\alpha, \beta)$, tenemos

$$\frac{\partial F}{\partial \alpha} = -2(\mu_Y - \beta\mu_X - \alpha), \qquad \frac{\partial F}{\partial \beta} = 2\beta\sigma_X^2 - 2\sigma_{XY} - 2\mu_X(\mu_Y - \beta\mu_X - \alpha)$$

Igualando estas ecuaciones a cero, que es la condición necesaria para que $F(\alpha, \beta)$ sea un mínimo, encontramos

$$\mu_Y = \alpha + \beta\mu_X, \qquad \beta\sigma_X^2 = \sigma_{XY}$$

Por consiguiente, si $y = \alpha + \beta x$, entonces $y - \mu_Y = \beta(x - \mu_X)$ o

$$y - \mu_Y = \frac{\sigma_{XY}}{\sigma_X^2}(x - \mu_X)$$

o

$$\frac{y - \mu_Y}{\sigma_Y} = \rho\left(\frac{x - \mu_X}{\sigma_X}\right)$$

Es importante observar la similitud de la demostración anterior para poblaciones, usando esperanzas, con la correspondiente demostración para muestras usando sumatorias. En general, los resultados para muestras tienen resultados análogos para poblaciones y viceversa.

8.40. La función de densidad conjunta de las variables aleatorias X y Y es

$$f(x, y) = \begin{cases} \frac{2}{3}(x + 2y) & 0 \leqq x \leqq 1, 0 \leqq y \leqq 1 \\ 0 & \text{de otra manera} \end{cases}$$

Encuentre la curva de regresión de mínimos cuadrados de (a) Y en X, (b) X en Y.

(a) La función de densidad marginal de X es

$$f_1(x) = \int_0^1 \frac{2}{3}(x + 2y)\, dy = \frac{2}{3}(x + 1)$$

para $0 \leqq x \leqq 1$, y $f_1(x) = 0$ de otra manera. Ya que, para $0 \leqq x \leqq 1$, la densidad condicional de Y dado X es

$$f_2(y|x) = \frac{f(x, y)}{f_1(x)} = \begin{cases} \frac{x + 2y}{x + 1} & 0 \leqq y \leqq 1 \\ 0 & y < 0 \text{ o } y > 1 \end{cases}$$

y la curva de regresión de mínimos cuadrados de Y en X está dada por

$$y = E(Y|X = x) = \int_{-\infty}^{\infty} y f_2(y|x) dy$$

$$= \int_0^1 y\left(\frac{x + 2y}{x + 1}\right) dy = \frac{3x + 4}{6x + 6}$$

Ni $f_2(y \mid x)$ ni la curva de regresión de mínimos cuadrados está definida cuando $x < 0$ o $x > 1$.

(b) Para $0 \leqq y \leqq 1$, la función de densidad marginal de Y es

$$f_2(y) = \int_0^1 \frac{2}{3}(x + 2y) dx = \frac{1}{3}(1 + 4y)$$

Por consiguiente, para $0 \leqq y \leqq 1$, la densidad condicional de X dado Y es

$$f_1(x|y) = \frac{f(x, y)}{f_2(y)} = \begin{cases} \frac{2x + 4y}{1 + 4y} & 0 \leqq x \leqq 1 \\ 0 & x < 0 \text{ o } x > 1 \end{cases}$$

y la curva de regresión de mínimos cuadrados de X en Y está dada por

$$x = E(X|Y = y) = \int_{-\infty}^{\infty} x f_1(x|y) dx$$

$$= \int_0^1 x\left(\frac{2x + 4y}{1 + 4y}\right) dx = \frac{2 + 6y}{3 + 12y}$$

Ni $f_1(x \mid y)$ ni la curva de regresión de mínimos cuadrados está definida cuando $y < 0$ o $y > 1$.

Observe que las dos curvas de regresión $y = (3x + 4)/(6x + 6)$ y $x = (2 + 6y)/(3 + 12y)$ son diferentes.

8.41. Encuentre (*a*) $\bar{X}$, (*b*) $\bar{Y}$, (*c*) σ_X^2, (*d*) σ_Y^2, (*e*) σ_{XY}, (*f*) ρ para la distribución del problema 8.40.

(*a*) $$\bar{X} = \int_{x=0}^{1}\int_{y=0}^{1} x\left[\frac{2}{3}(x+2y)\right]dx\,dy = \frac{5}{9}$$

(*b*) $$\bar{Y} = \int_{x=0}^{1}\int_{y=0}^{1} y\left[\frac{2}{3}(x+2y)\right]dx\,dy = \frac{11}{18}$$

(*c*) $$\overline{X^2} = \int_{x=0}^{1}\int_{y=0}^{1} x^2\left[\frac{2}{3}(x+2y)\right]dx\,dy = \frac{7}{18}$$

Entonces $$\sigma_X^2 = \overline{X^2} - \bar{X}^2 = \frac{7}{18} - \left(\frac{5}{9}\right)^2 = \frac{13}{162}$$

(*d*) $$\overline{Y^2} = \int_{x=0}^{1}\int_{y=0}^{1} y^2\left[\frac{2}{3}(x+2y)\right]dx\,dy = \frac{4}{9}$$

Entonces $$\sigma_Y^2 = \overline{Y^2} - \bar{Y}^2 = \frac{4}{9} - \left(\frac{11}{18}\right)^2 = \frac{23}{324}$$

(*e*) $$\overline{XY} = \int_{x=0}^{1}\int_{y=0}^{1} xy\left[\frac{2}{3}(x+2y)\right]dx\,dy = \frac{1}{3}$$

Entonces $$\sigma_{XY} = \overline{XY} - \bar{X}\bar{Y} = \frac{1}{3} - \left(\frac{5}{9}\right)\left(\frac{11}{18}\right) = -\frac{1}{162}$$

(*f*) $$\rho = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} = \frac{-1/162}{\sqrt{13/162}\sqrt{23/324}} = -0.0818$$

Observe que el coeficiente de correlación lineal es pequeño. Esto es de esperarse a partir de la observación de las rectas de regresión de mínimos cuadrados obtenidas en el problema 8.42.

8.42. Escriba las rectas de regresión de mínimos cuadrados de (*a*) Y en X, (*b*) X en Y para el problema 8.40.

(*a*) La recta de regresión de Y en X es $\dfrac{y-\bar{Y}}{\sigma_Y} = \rho\left(\dfrac{x-\bar{X}}{\sigma_X}\right)$ o

$$y - \bar{Y} = \frac{\sigma_{XY}}{\sigma_X^2}(x-\bar{X}) \qquad \text{o} \qquad y - \frac{11}{18} = \frac{-1/162}{13/162}\left(x - \frac{5}{9}\right)$$

(*b*) La recta de regresión de X en Y es $\dfrac{x-\bar{X}}{\sigma_X} = \rho\left(\dfrac{y-\bar{Y}}{\sigma_Y}\right)$ o

$$x - \bar{X} = \frac{\sigma_{XY}}{\sigma_Y^2}(y-\bar{Y}) \qquad \text{o} \qquad x - \frac{5}{9} = \frac{-1/162}{23/324}\left(y - \frac{11}{18}\right)$$

TEORÍA MUESTRAL DE LA REGRESIÓN

8.43. En el problema 8.11 encontramos que la ecuación de regresión de y en x es $y = 35.82 + 0.476x$. Al nivel de significancia de 0.05 pruebe la hipótesis de que el coeficiente de regresión de la ecuación de regresión de la población es tan pequeño como 0.108.

$$t = \frac{\beta - b}{s_{y.x}/s_x}\sqrt{n-2} = \frac{0.476 - 0.180}{1.28/2.66}\sqrt{12-2} = 1.95$$

Dado que $s_{y.x} = 1.28$ (calculado en el problema 8.22) y $s_x = \sqrt{\overline{x^2} - \bar{x}^2} = 2.66$ del problema 8.11.

Con base en una prueba de una cola de la distribución de Student al nivel de 0.05, podemos rechazar la hipótesis de que el coeficiente de regresión es tan pequeño como 0.180 si $t > t_{0.95} = 1.81$ para $12 - 2 = 10$ grados de libertad. Por consiguiente, podemos rechazar la hipótesis.

8.44. Encuentre los límites de confianza del 95% para coeficiente de regresión del problema 8.43.

$$\beta = b + \frac{t}{\sqrt{n-2}}\left(\frac{s_{y.x}}{s_x}\right)$$

Entonces los límites de confianza del 95% para β (obtenidos remplazando $t = \pm t_{0,975} = \pm 2.23$ para $12 - 2 = 10$ grados de libertad) están dados por

$$b \pm \frac{2.23}{\sqrt{12-2}}\left(\frac{s_{y.x}}{s_x}\right) = 0.476 \pm \frac{2.23}{\sqrt{10}}\left(\frac{1.28}{2.66}\right) = 0.476 \pm 0.340$$

es decir, tenemos confianza del 95% de que β está entre 0.136 y 0.816.

8.45. En el problema 8.11, encuentre los límites de confianza del 95% para las estaturas de los hijos cuyos padres tienen estaturas de (*a*) 65.0, (*b*) 70.0 pulgadas.

Dado que $t_{0.975} = 2.23$ para $12 - 2 = 10$ grados de libertad, los límites de confianza del 95% para y_p son

$$y_0 \pm \frac{2.23}{\sqrt{n-2}} s_{y.x}\sqrt{n+1+\frac{n(x_0-\bar{x})^2}{s_x^2}}$$

donde $y_0 = 35.82 + 0.476x_0$ (problema 8.11), $s_{y.x} = 1.28$, $s_x = 2.66$ (problema 8.43) y $n = 12$.

(*a*) Si $x_0 = 65.0$, $y_0 = 66.76$ pulgadas. También $(x_0 - \bar{x})^2 = (65.0 - 800/12)^2 = 2.78$. Entonces, los límites de confianza del 95% son

$$66.76 \pm \frac{2.23}{\sqrt{10}}(1.28)\sqrt{12+1+\frac{12(2.78)}{(2.66)^2}} = 66.76 \pm 3.80 \text{ pulgadas}$$

es decir, podemos tener confianza de cerca del 95% de que las estaturas de los hijos están entre 63.0 y 70.6 pulgadas.

(*b*) Si $x_0 = 70.0$, $y_0 = 69.14$. También $(x_0 - \bar{x})^2 = (70.0 - 800/12)^2 = 11.11$. Entonces, se calcula que los límites de confianza del 95% son 69.14 ± 5.09 pulgadas, o sea, podemos tener el 95% de confianza de que la estatura de los hijos estará entre 64.1 y 74.2 pulgadas.

Observe que para valores grandes de n, los límites de confianza del 95% están dados por aproximadamente $y_0 \pm 1.96 s_{y.x}$ o $y_0 \pm 2 s_{y.x}$ siempre y cuando $x_0 - \bar{x}$ no sea demasiado grande. Esto concuerda con los resultados aproximados mencionados en la página 289. Los métodos en este problema son válidos sin tener el cuenta el tamaño de n o $x_0 - \bar{x}$, es decir, los métodos muestrales son exactos para una población normal.

8.46. En el problema 8.11, encuentre los límites de confianza del 95% para la estatura media de los hijos cuyos padres tengan estaturas de (*a*) 65.0, (*b*) 70.0 pulgadas.

Dado que $t_{0.975} = 2.23$ para 10 grados de libertad, los límites de confianza del 95% para $\bar{y}_p$ son

$$y_0 \pm \frac{2.23}{\sqrt{10}} s_{y.x}\sqrt{1+\frac{(x_0-\bar{x})^2}{s_x^2}}$$

donde $y_0 = 35.82 + 0.476x_0$ (problema 8.11), $s_{y.x} = 1.28$ (problema 8.43).

(*a*) Si $x_0 = 65.0$, encontramos [compare con el problema 8.45(*a*)] que los límites de confianza del 95% son 66.76 ± 1.07 pulgadas, es decir, podemos tener confianza del 95% de que la estatura media de todos los hijos cuyos padres tengan estatura de 65.0 pulgadas estará entre 65.7 y 67.8 pulgadas.

(*b*) Si $x_0 = 70.0$, encontramos [compare con el problema 8.45(*b*)] que los límites de confianza del 95% son 69.14 ± 1.45 pulgadas, es decir, podemos tener confianza del 95% de que la estatura media de todos los hijos cuyos padres tengan estatura de 70.0 pulgadas estará entre 67.7 y 70.6 pulgadas.

TEORÍA MUESTRAL DE LA CORRELACIÓN

8.47. Se calculó que un coeficiente de correlación basado en una muestra de tamaño 18, era 0.32. ¿Podemos concluir con un nivel de significancia de (*a*) 0.05, (*b*) 0.01 que el correspondiente coeficiente de correlación de la población es significativamente mayor que cero?

Vamos a escoger entre las hipótesis (H_0: $\rho = 0$) y (H_1: $\rho > 0$).

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.32\sqrt{18-2}}{\sqrt{1-(0.32)^2}} = 1.35$$

(*a*) Con base en una prueba de una cola de la distribución de Student a un nivel de 0.05, podemos rechazar la hipótesis H_0 si $t > t_{0.95} = 1.75$ para $18 - 2 = 16$ grados de libertad. Por consiguiente, no podemos rechazar H_0 al nivel de 0.05.

(*b*) Dado que no podemos rechazar H_0 al nivel de 0.05, ciertamente no podemos rechazarla al nivel de 0.01.

8.48. ¿Cuál es el tamaño mínimo necesario de la muestra para que podamos concluir a un nivel de 0.05 que un coeficiente de correlación de 0.32 es significativamente mayor que cero?

Al nivel de 0.05 y usando una prueba de una cola de la distribución de Student, el valor mínimo de n debe ser tal que

$$\frac{0.32\sqrt{n-2}}{\sqrt{1-(0.32)^2}} = t_{0.95} \text{ para } n-2 \text{ grados de libertad}$$

Para $n = 26$, $\nu = 24$, $t_{0.95} = 1.71$, $t = 0.32\sqrt{24}/\sqrt{1-(0.32)^2} = 1.65$.

Para $n = 27$, $\nu = 25$, $t_{0.95} = 1.71$, $t = 0.32\sqrt{25}/\sqrt{1-(0.32)^2} = 1.69$.

Para $n = 28$, $\nu = 26$, $t_{0.95} = 1.71$, $t = 0.32\sqrt{26}/\sqrt{1-(0.32)^2} = 1.72$.

Entonces el tamaño mínimo de la muestra es n = 28.

8.49. Se calculó que un coeficiente de correlación basado en una muestra de tamaño 24, es $r = 0.75$. ¿Podemos rechazar la hipótesis al nivel de significancia de 0.05 de que el coeficiente de correlación de la población es tan pequeño como (*a*) $\rho = 0.60$, (*b*) $\rho = 0.50$?

(*a*) $$Z = 1.1513 \log\left(\frac{1+0.75}{1-0.75}\right) = 0.9730, \qquad \mu_Z = 1.1513 \log\left(\frac{1+0.60}{1-0.60}\right) = 0.6932,$$

$$\sigma_Z = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{21}} = 0.2182$$

La variable estandarizada es entonces

$$z = \frac{Z-\mu_Z}{\sigma_Z} = \frac{0.9730-0.6932}{0.2182} = 1.28$$

Al nivel de significancia de 0.05 usando una prueba de una cola de la distribución normal, podemos rechazar la hipótesis sólo si z fuera mayor que 1.64. Por consiguiente, no podemos rechazar la hipótesis de que el coeficiente de correlación de la población es tan pequeño como 0.60.

(*b*) Si $\rho = 0.50$, $\mu_z = 1.1513$, log 3 = 0.5493 y $z = (0.9730 = 0.5493)/0.2182 = 1.94$. Por consiguiente, podemos rechazar la hipótesis al nivel de significancia de 0.05 de que el coeficiente de correlación de la población es tan pequeño como $\rho = 0.50$.

8.50. Se calculó que el coeficiente de correlación entre las notas finales de física y matemáticas para un grupo de 21 estudiantes fue de 0.80. Encuentre los límites de confianza del 95% para este coeficiente.

Dado que $r = 0.80$ y $n = 21$, los límites de confianza del 95% para μ_z están dados por

$$Z \pm 1.96\sigma_Z = 1.1513 \log\left(\frac{1+r}{1-r}\right) \pm 1.96\left(\frac{1}{\sqrt{n-3}}\right) = 1.0986 \pm 0.4620$$

Entonces μ_z tiene el intervalo de confianza de 0.5366 a 1.5606.

Si $$\mu_Z = 1.1513 \log\left(\frac{1+\rho}{1-\rho}\right) = 0.5366, \qquad \rho = 0.4904.$$

Si $$\mu_Z = 1.1513 \log\left(\frac{1+\rho}{1-\rho}\right) = 1.5606, \qquad \rho = 0.9155.$$

Por lo tanto, los límites de confianza del 95% para ρ son 0.49 y 0.92.

8.51. Se obtienen dos coeficientes de correlación de muestras de tamaño $n_1 = 28$ y $n_2 = 35$ donde se calculó que $r_1 = 0.50$ y $r_2 = 0.30$, respectivamente. ¿Hay alguna diferencia significativa al nivel de 0.05 entre los dos coeficientes?

$$Z_1 = 1.1513 \log\left(\frac{1+r_1}{1-r_1}\right) = 0.5493, \qquad Z_2 = 1.1513 \log\left(\frac{1+r_2}{1-r_2}\right) = 0.3095$$

y $$\sigma_{Z_1-Z_2} = \sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}} = 0.2669$$

Queremos decidir entre las hipótesis (H_0: $\mu_{z_1} = \mu_{z_2}$) y (H_1: $\mu_{z_1} \neq \mu_{z_2}$).
Bajo la hipótesis H_0,

$$z = \frac{Z_1 - Z_2 - (\mu_{Z_1} - \mu_{Z_2})}{\sigma_{Z_1-Z_2}} = \frac{0.5493 - 0.3095 - 0}{0.2669} = 0.8985$$

Usando una prueba de dos colas de la distribución normal, podemos rechazar H_0 sólo si $z > 1.96$ o $z < -1.96$. Por consiguiente, no podemos rechazar H_0, y concluimos que los resultados no son significativamente diferentes al nivel de 0.05.

PROBLEMAS DIVERSOS

8.52. Demuestre la fórmula (*25*), página 290.

Para la recta de mínimos cuadrados tenemos, a partir de los problemas 8.20 y 8.21,

$$s_{y.x}^2 = \frac{\sum(y-\bar{y})^2}{n} - b\frac{\sum(x-\bar{x})(y-\bar{y})}{n}$$

Pero por definición

$$\frac{\sum(y-\bar{y})^2}{n} = s_y^2 \qquad \frac{\sum(x-\bar{x})(y-\bar{y})}{n} = s_{xy}$$

y según (*6*) en la página 287,

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{s_{xy}}{s_x^2}$$

De ahí, $$s_{y.x}^2 = s_y^2 - \frac{s_{xy}^2}{s_x^2} = s_y^2\left[1 - \left(\frac{s_{xy}}{s_x s_y}\right)^2\right] = s_y^2(1-r^2)$$

Una fórmula análoga es válida para la población (*véase* el problema 8.54).

8.53. Demuestre que $E[(Y-\bar{Y})^2] = E[(Y-Y_{est})^2] + E[(Y_{est}-\bar{Y})^2]$ para el caso de (*a*) una recta de mínimos cuadrados, (*b*) una parábola de mínimos cuadrados.

Tenemos $$Y - \bar{Y} = (Y - Y_{est}) + (Y_{est} - \bar{Y})$$

Entonces $$(Y-\bar{Y})^2 = (Y-Y_{est})^2 + (Y_{est}-\bar{Y})^2 + 2(Y-Y_{est})(Y_{est}-\bar{Y})$$

y así $$E[(Y-\bar{Y})^2] = E[(Y-Y_{est})^2] + E[(Y_{est}-\bar{Y})^2] + 2E[(Y-Y_{est})(Y_{est}-\bar{Y})]$$

El resultado requerido se deduce si podemos demostrar que el último término es cero.

(*a*) Para la regresión lineal, $Y_{est} = \alpha + \beta X$. Entonces

$$\begin{aligned} E[(Y-Y_{est})(Y_{est}-\bar{Y})] &= E[(Y-\alpha-\beta X)(\alpha+\beta X-\bar{Y})] \\ &= (\alpha-\bar{Y})E(Y-\alpha-\beta X) + \beta E(XY-\alpha X-\beta X^2) \\ &= 0 \end{aligned}$$

debido a las ecuaciones normales

$$E(Y-\alpha-\beta X) = 0, \qquad E(XY-\alpha X-\beta X^2) = 0$$

(compare con el problema 8.3).

(*b*) Para la regresión parabólica, $Y_{est} = \alpha + \beta X + \gamma X^2$. Entonces

$$\begin{aligned} E[(Y-Y_{est})(Y_{est}-\bar{Y})] &= E[(Y-\alpha-\beta X-\gamma X^2)(\alpha+\beta X+\gamma X^2-\bar{Y})] \\ &= (\alpha-\bar{Y})E(Y-\alpha-\beta X-\gamma X^2) + \beta E[X(Y-\alpha-\beta X-\gamma X^2)] \\ &\quad + \gamma E[X^2(Y-\alpha-\beta X-\gamma X^2)] \\ &= 0 \end{aligned}$$

debido a las ecuaciones normales

$$E(Y-\alpha-\beta X-\gamma X^2) = 0, \qquad E[X(Y-\alpha-\beta X-\gamma X^2)] = 0, \qquad E[X^2(Y-\alpha-\beta X-\gamma X^2)] = 0$$

Compare las ecuaciones (*19*), página 289.

El resultado puede extenderse a curvas de mínimos cuadrados de órdenes superiores.

8.54. Demuestre que $\sigma_{Y.X}^2 = \sigma_Y^2(1-\rho^2)$ para la regresión de mínimos cuadrados.

Por definición del coeficiente de correlación generalizado ρ junto con el problema 8.53, tenemos para cada uno de los casos lineal o parabólico,

$$\rho^2 = \frac{E[(Y_{est}-\bar{Y})^2]}{E[(Y-\bar{Y})^2]} = 1 - \frac{E[(Y-Y_{est})^2]}{E[(Y-\bar{Y})^2]} = 1 - \frac{\sigma_{Y.X}^2}{\sigma_Y^2}$$

y el resultado se deduce inmediatamente.

El resultado puede extenderse a curvas de mínimos cuadrados de órdenes superiores.

8.55. Demuestre que para el caso de la regresión lineal el coeficiente de correlación como se definió en (*45*) se reduce a lo definido en (*40*).

En el caso de la regresión lineal, el cuadrado del coeficiente de correlación, es decir, el *coeficiente de determinación*, como se dio en (*45*) está dado por

(*1*) $$\rho^2 = \frac{E[(Y_{est}-\bar{Y})^2]}{E[(Y-\bar{Y})^2]} = \frac{E[(\alpha+\beta X-\bar{Y})^2]}{\sigma_Y^2}$$

Pero dado que $\bar{Y} = \alpha + \beta\bar{X}$,

$$(2) \qquad E[(\alpha + \beta X - \bar{Y})^2] = E[\beta^2(X - \bar{X})^2] = \beta^2 E[(X - \bar{X})^2]$$
$$= \frac{\sigma_{XY}^2}{\sigma_X^4}\sigma_X^2 = \frac{\sigma_{XY}^2}{\sigma_X^2}$$

Entonces (*1*) se convierte en

$$(3) \qquad \rho^2 = \frac{\sigma_{XY}^2}{\sigma_X^2\sigma_Y^2} \qquad \text{o} \qquad \rho = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

como se requirió para la demostración (el signo correcto para ρ se incluye en σ_{XY}^2).

8.56. Refiérase a la tabla 8-21. (*a*) Encuentre una parábola de regresión de mínimos cuadrados que se ajuste a los datos. (*b*) Calcule los valores de regresión (denominados *valores de tendencia*) para los años dados y compárelos con los valores reales. (*c*) Estime la población en 1945. (*d*) Estime la población en 1960 y compárela con el valor real, 179.3. (*e*) Estime la población en 1840 y compárela con el valor real, 17.1.

Tabla 8-21

Año	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950
Población de E.U.A. (en millones)	23.2	31.4	39.8	50.2	62.9	76.0	92.0	105.7	122.8	131.7	151.1

Fuente: Bureau of the census.

(*a*) Sean x y y las variables que denotan, respectivamente, el año y la población durante ese año. La ecuación de la parábola de mínimos cuadrados que se ajusta a los datos es

$$(1) \qquad y = a + bx + cx^2$$

donde a, b y c se encontraron a partir de las ecuaciones normales

$$(2) \qquad \begin{aligned} \textstyle\sum y &= an + b\sum x + c\sum x^2 \\ \textstyle\sum xy &= a\sum x + b\sum x^2 - c\sum x^3 \\ \textstyle\sum x^2y &= a\sum x^2 + b\sum x^3 + c\sum x^4 \end{aligned}$$

Es conveniente localizar el origen de manera que el año de la mitad, 1900, corresponda a $x = 0$, y escoger una unidad que haga que los años 1910, 1020, 1930, 1040, 1950 y 1890, 1880, 1870, 1860, 1850 correspondan a 1, 2, 3, 4, 5 y $-1, -2, -3, -4, -5$, respectivamente. Con esta selección $\sum x$ y $\sum x^3$ son cero y las ecuaciones de (*2*) están simplificadas.

El trabajo involucrado en el cálculo puede disponerse como en la tabla 8-22. Las ecuaciones normales (*2*) se convierten en

$$(3) \qquad \begin{aligned} 11a + 110c &= 886.8 \\ 110b &= 1429.8 \\ 110a + 1958c &= 9209.0 \end{aligned}$$

A partir de la segunda ecuación en (*3*), $b = 13.00$; a partir de la primera y la tercera ecuación, $a = 76.64$, $c = 0.3974$. Entonces, la ecuación requerida es

$$(4) \qquad y = 76.64 + 13.00x + 0.3974x^2$$

donde el origen, $x = 0$, es julio 1 de 1990, y la unidad de x es 10 años.

(*b*) Los valores de tendencia, obtenidos al sustituir $x = -5, -4, -3, -3, -1, 0, 1, 2, 3, 4, 5$ en (*4*), se muestran en la tabla 8-23 junto con los valores reales. Se ve que la concordancia es buena.

(*c*) 1945 corresponde a $x = 4.5$, para lo cual $y = 76.64 + 13.00(4.5) + 0.3974(4.5)^2 = 143.2$.

(*d*) 1960 corresponde a $x = 6$, para lo cual $y = 76.64 + 13.00(6) + 0.3974(6)^2 = 168.9$. Esto no coincide muy bien con el valor real, 179.3.

(*e*) 1840 corresponde a $x = -6$, para lo cual $y = 76.64 + 13.00(-6) + 0.3974(-6)^2 = 12.9$. Esto no coincide con el valor real, 17.1.

Este ejemplo ilustra el hecho de que una relación que parece ser satisfactoria para un rango de valores, no necesariamente es satisfactoria para una rango extendido de valores.

Tabla 8-22

Año	x	y	x^2	x^3	x^4	xy	x^2y
1850	−5	23.2	25	−125	625	−116.0	580.0
1860	−4	31.4	16	−64	256	−125.6	502.4
1870	−3	39.8	9	−27	81	−119.4	358.2
1880	−2	50.2	4	−8	16	−100.4	200.8
1890	−1	62.9	1	−1	1	−62.9	62.9
1900	0	76.0	0	0	0	0	0
1910	1	92.0	1	1	1	92.0	92.0
1920	2	105.7	4	8	16	211.4	422.8
1930	3	122.8	9	27	81	368.4	1105.2
1940	4	131.7	16	64	256	526.8	2107.2
1950	5	151.1	25	125	625	755.5	3777.5
	$\sum x = 0$	$\sum y =$ 886.8	$\sum x^2 =$ 110	$\sum x^3 = 0$	$\sum x^4 =$ 1958	$\sum xy =$ 1429.8	$\sum x^2y =$ 9209.0

Tabla 8-23

Año	$x = -5$ 1850	$x = -4$ 1860	$x = -3$ 1870	$x = -2$ 1880	$x = -1$ 1890	$x = 0$ 1900	$x = 1$ 1910	$x = 2$ 1920	$x = 3$ 1930	$x = 4$ 1940	$x = 5$ 1950
Valor de tendencia	21.6	31.0	41.2	52.2	64.0	76.6	90.0	104.2	119.2	135.0	151.6
Valor real	23.2	31.4	39.8	50.2	62.9	76.0	92.0	105.7	122.8	131.7	151.1

8.57. Los precios promedio de las acciones y los bonos presentados en la Bolsa de Valores de Nueva York durante los años 1950 a 1959 están dados en la tabla 8-24. (*a*) Encuentre el coeficiente de correlación, (*b*) interprete los resultados.

Tabla 8-24

Año	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959
Precio promedio de las acciones (dólares)	35.22	39.87	41.85	43.23	40.06	53.29	54.14	49.12	40.71	55.15
Precio promedio de los bonos (dólares)	102.43	100.93	97.43	97.81	98.32	100.07	97.08	91.59	94.85	94.65

Fuente: Bolsa de Valores de Nueva York.

(*a*) Sean x y y los precios promedio de las acciones y los bonos, los cálculos de los coeficientes de correlación pueden organizarse como en la tabla 8-25. Observe que el año solamente se usa para especificar los valores correspondientes de x y y.

Tabla 8-25

x	y	$x' = x - \bar{x}$	$y' = y - \bar{y}$	x'^2	$x'y'$	y'^2
35.22	102.43	−10.04	4.91	100.80	−49.30	24.11
39.87	100.93	−5.39	3.41	29.05	−18.38	11.63
41.85	97.43	−3.41	−0.09	11.63	0.31	0.01
43.23	97.81	−2.03	0.29	4.12	−0.59	0.08
40.06	98.32	−5.20	0.80	27.04	−4.16	0.64
53.29	100.07	8.03	2.55	64.48	20.48	6.50
54.14	97.08	8.88	−0.44	78.85	−3.91	0.19
49.12	91.59	3.86	−5.93	14.90	−22.89	35.16
40.71	94.85	−4.55	−2.67	20.70	12.15	7.13
55.15	94.65	9.89	−2.87	97.81	−28.38	8.24
$\sum x =$ 452.64	$\sum x =$ 975.16			$\sum x'^2 =$ 449.38	$\sum x'y' =$ −94.67	$\sum y'^2 =$ 93.69
$\bar{x} =$ 45.26	$\bar{y} =$ 97.52					

Entonces por la fórmula de producto-momento,

$$r = \frac{\sum x'y'}{\sqrt{(\sum x'^2)(\sum y'^2)}} = \frac{-94.67}{\sqrt{(449.38)(93.69)}} = -0.4614$$

(*b*) Concluimos que hay cierta correlación negativa entre el precio de las acciones y de los bonos (o sea, existe la tendencia de que los precios de las acciones bajen cuando el de los bonos sube, y viceversa), aunque esta relación no es marcada.

Otro método.

La tabla 8-26 muestra los rangos de los precios promedio de las acciones para los años 1950 a 1959 en orden ascendente. También muestra las diferencias en rango d y $\sum d^2$.

Tabla 8-26

Año	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	
Precios de las acciones en orden de rango	1	2	5	6	3	8	9	7	4	10	
Precios de los bonos en orden de rango	10	9	5	6	7	8	4	1	3	2	
Diferencias en el rango, d	−9	−7	0	0	−4	0	5	6	1	8	
d^2	81	49	0	0	16	0	25	36	1	64	$\sum d^2 =$ 272

Entonces
$$r_{\text{rank}} = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6(272)}{10(10^2-1)} = -0.6485$$

Esto se compara favorablemente con el resultado del primer método.

8.58. La tabla 8-27 muestra las distribuciones de frecuencia de las notas finales de matemáticas y física de 100 estudiantes. Según esta tabla determine (*a*) el número de estudiantes que recibieron notas entre 70 a 79 en matemáticas y entre 80 a 89 en física, (*b*) el porcentaje de estudiantes con notas de matemáticas inferiores a 70, (*c*) el número de estudiantes que obtuvieron notas de 70 o más en física y menos de 80 en matemáticas, (*d*) el porcentaje de estudiantes que pasaron al menos una de las materias, suponiendo que 60 es la nota mínima para pasar el curso.

Tabla 8-27
NOTAS DE MATEMÁTICAS

NOTAS DE FÍSICA	40–49	50–59	60–69	70–79	80–89	90–99	TOTALES
90–99				2	4	4	10
80–89			1	4	6	5	16
70–79			5	10	8	1	24
60–69	1	4	9	5	2		21
50–59	3	6	6	2			17
40–49	3	5	4				12
TOTALES	7	15	25	23	20	10	100

(*a*) Desplácese hacia abajo por la columna titulada 70−90 (notas de matemáticas) hasta la fila marcada como 80-90 (notas de física). El 4 de intersección da el número requerido de estudiantes.

(*b*) El número total de estudiantes con notas de matemáticas por debajo de 70

= (número con notas 40−49) + (número con notas 50−59) + (número con notas 60−69)

= 7 + 15 + 25 = 47

Porcentaje de estudiantes con notas de matemáticas inferiores a 70 = 47/100 = 47%.

(*c*) El número requerido de estudiantes en el total de las intersecciones de la tabla 8-28, que representan parte de la tabla 8-27.

Número requerido de estudiantes = 1 + 5 + 2 + 4 + 10 = 22.

Tabla 8-28
NOTAS DE MATEMÁTICAS

NOTAS DE FÍSICA	60–69	70–79
90–99		2
80–89	1	4
70–79	5	10

Tabla 8-29
NOTAS DE MATEMÁTICAS

NOTAS DE FÍSICA	40–49	50–59
50–59	3	6
40–49	3	5

(*d*) En relación con la tabla 8-29, la cual se tomó de la tabla 8-27, se ve que el número de estudiantes con notas inferiores a 60 tanto en matemáticas como en física es 3 + 3 + 6 + 5 = 17. Entonces el número de estudiantes con notas de 60 o superiores tanto en física como en matemáticas o ambas es 100 − 17 = 83, y el porcentaje requerido es 83/100 = 83%.

La tabla 8-27 recibe el nombre de *tabla de frecuencia bidimensional* o *distribución de frecuencia bidimensional*. Cada cuadrado de la tabla se denomina *celda* y corresponde a un par de clases o intervalos de clase. El número indicado en la celda se llama *frecuencia de celda*. Por ejemplo, en la parte (*a*) el número 4 es la frecuencia de la celda correspondiente al par de los intervalos de clase 70-79 en matemáticas y 80-89 en física.

Los totales indicados en el último renglón y en la última columna se denominan *totales marginales* o *frecuencias marginales*. Ellos corresponden, respectivamente, a las frecuencias de clase de las distribuciones de frecuencia de las notas de matemáticas y física.

8.59. Muestre de qué manera puede usted modificar la fórmula del problema 8.31 para el caso de datos agrupados como en la tabla 8-27.

Para datos agrupados podemos considerar los diferentes valores de las variables x y y a medida que coinciden con las marcas de clase, mientras f_x y f_y son las frecuencias de clase correspondientes o frecuencias marginales indicadas en la última fila de la tabla de frecuencia bidimensional. Si f representa las diferentes frecuencias de celda correspondientes a los pares de marcas de clase (x, y), entonces podemos remplazar la fórmula del problema8.31 por

$$r = \frac{n\sum fxy - (\sum f_x x)(\sum f_y y)}{\sqrt{[n\sum f_x x^2 - (\sum f_x x)^2][n\sum f_y y^2 - (\sum f_y y)^2]}} \tag{1}$$

Si $x = x_0 + c_x u_x$ y $y = y_0 + c_y u_y$, donde c_x y c_y son las amplitudes de los intervalos de clase (que suponemos constantes) y x_0 y y_0 son marcas de clase arbitrarias correspondientes a las variables, la fórmula anterior se convierte en

$$r = \frac{n\sum fu_x u_y - (\sum f_x u_x)(\sum f_y u_y)}{\sqrt{[n\sum f_x u_x^2 - (\sum f_x u_x)^2][n\sum f_y u_y^2 - (\sum f_y u_y)^2]}} \tag{2}$$

Éste es el *método clave* usado en el capítulo 5 como método abreviado para calcular medias, desviaciones estándar y momentos superiores.

8.60. Encuentre el coeficiente de correlación lineal de las notas de matemáticas y física del problema 8.58.

Usamos la fórmula (2) del problema 8.59. El trabajo se puede disponer como en la tabla 8-30, la cual se denomina *tabla de correlación*.

Tabla 8-30

		Notas de Matemáticas, x									
		x	44.5	54.5	64.5	74.5	84.5	94.5	f_y	$f_y u_y$	$f_y u_y^2$	Suma de los números de las esquinas de cada fila
Notas de física, y	y	u_y \ u_x	−2	−1	0	1	2	3				
	94.5	2				2 [4]	4 [16]	4 [24]	10	20	40	44
	84.5	1			1 [0]	4 [4]	6 [12]	5 [15]	16	16	16	31
	74.5	0			5 [0]	10 [0]	8 [0]	1 [0]	24	0	0	0
	64.5	−1	1 [2]	4 [4]	9 [0]	5 [−5]	2 [−4]		21	−21	21	−3
	54.5	−2	3 [12]	6 [12]	6 [0]	2 [−4]			17	−34	68	20
	44.5	−3	3 [18]	5 [15]	4 [0]				12	−36	108	33
	f_x		7	15	25	23	20	10	$\sum f_x = \sum f_y = n = 100$	$\sum f_y u_y = -55$	$\sum f_y u_y^2 = 253$	$\sum fu_x u_y = 125$
	$f_x u_x$		−14	−15	0	23	40	30	$\sum f_x u_x = 64$			
	$f_x u_x^2$		28	15	0	23	80	90	$\sum f_x u_x^2 = 236$			
	Suma de los números de las esquinas en cada columna		32	31	0	−1	24	39	$\sum fu_x u_y = 125$			

Verificación

El número de la esquina de cada celda representa el producto fu_xu_y, donde f es la frecuencia de celda. La suma de los números de las esquinas en cada fila se indica en la fila correspondiente de la última columna. La suma de los números de las esquinas en cada columna se indica en la columna correspondiente en la última fila. Los totales finales de la última fila y de la última columna son iguales y representan $\sum fu_xu_y$.

De la tabla 8-30 tenemos,

$$r = \frac{n\sum fu_xu_y - (\sum f_xu_x)(\sum f_yu_y)}{\sqrt{[n\sum f_xu_x^2 - (\sum f_xu_x)^2][n\sum f_yu_y^2 - (\sum f_yu_y)^2]}}$$

$$= \frac{(100)(125) - (64)(-55)}{\sqrt{[(100)(236) - (64)^2][(100)(253) - (-55)^2]}} = \frac{16{,}020}{\sqrt{(19{,}504)(22{,}275)}} = 0.7686$$

8.61. Use la tabla de correlación del problema 8.60 para calcular (*a*) S_x, (*b*) S_y, (*c*) S_{xy} y verificar la fórmula $r = S_{xy}/S_xS_y$.

(*a*) $$s_x = c_x\sqrt{\frac{\sum f_xu_x^2}{n} - \left(\frac{\sum f_xu_x}{n}\right)^2} = 10\sqrt{\frac{236}{100} - \left(\frac{64}{100}\right)^2} = 13.966$$

(*b*) $$s_y = c_y\sqrt{\frac{\sum f_yu_y^2}{n} - \left(\frac{\sum f_yu_y}{n}\right)^2} = 10\sqrt{\frac{253}{100} - \left(\frac{-55}{100}\right)^2} = 14.925$$

(*c*) $$s_{xy} = c_xc_y\left[\frac{\sum fu_xu_y}{n} - \left(\frac{\sum f_xu_x}{n}\right)\left(\frac{\sum f_yu_y}{n}\right)\right] = (10)(10)\left[\frac{125}{100} - \left(\frac{64}{100}\right)\left(\frac{-55}{100}\right)\right] = 160.20$$

Por consiguiente, las desviaciones estándar de las notas de matemáticas y física son 14.0 y 14.9, respectivamente, mientras que su covarianza es 160.2. Tenemos

$$\frac{s_{xy}}{s_xs_y} = \frac{160.20}{(13.966)(14.925)} = 0.7686$$

lo que concuerda con r tal como se encontró en el problema 8.60.

8.62. Escriba las ecuaciones de las rectas de regresión de (*a*) y en x, (*b*) x en y para los datos del problema 8.60.

De la tabla 8-30 tenemos

$$\bar{x} = x_0 + c_x\frac{\sum f_xu_x}{n} = 64.5 + \frac{(10)(64)}{100} = 70.9$$

$$\bar{y} = y_0 + c_y\frac{\sum f_yu_y}{n} = 74.5 + \frac{(10)(-55)}{100} = 69.0$$

A partir de los resultados del problema 8.61, $s_x = 13.966$, $s_y = 14.925$ y $r = 0.7686$. Ahora usamos (*16*), página 288, para obtener las ecuaciones de las rectas de regresión.

(*a*) $$y - \bar{y} = \frac{rs_y}{s_x}(x - \bar{x}), \quad y - 69.0 = \frac{(0.7686)(14.925)}{13.966}(x - 70.9),$$

o $$y - 69.0 = 0.821(x - 70.9)$$

(*b*) $$x - \bar{x} = \frac{rs_x}{s_y}(y - \bar{y}), \quad y - 70.0 = \frac{(0.7686)(13.966)}{14.925}(y - 69.0),$$

o $$x - 70.9 = 0.719(y - 69.0)$$

8.63. Calcule los errores estándar del estimado (*a*) $S_{y.x}$, (*b*) $S_{x.y}$ para los datos del problema 8.60. Use los resultados del problema 8.61.

(*a*) $$s_{y.x} = s_y\sqrt{1 - r^2} = 14.925\sqrt{1 - (0.7686)^2} = 9.548$$

(*b*) $$s_{x.y} = s_x\sqrt{1 - r^2} = 13.966\sqrt{1 - (0.7686)^2} = 8.934$$

Problemas suplementarios

RECTA DE MÍNIMOS CUADRADOS

8.64. Ajuste una recta de mínimos cuadrados a los datos de la tabla 8-31 usando (*a*) x como variable independiente, (*b*) x como variable dependiente. Dibuje los datos y las rectas de mínimos cuadrados usando los mismos ejes de coordenadas.

Tabla 8-31

x	3	5	6	8	9	11
y	2	3	4	6	5	8

8.65. Para los datos del problema 8.64 encuentre (*a*) los valores de y cuando $x = 5$ y $x = 12$, (*b*) el valor de x cuando $y = 7$.

8.66. La tabla 8-32 muestra las notas finales de álgebra y física obtenidas por 10 estudiantes seleccionados al azar de un grupo grande de estudiantes. (*a*) Dibuje los datos. (*b*) Encuentre la recta de mínimos cuadrados que se ajuste a los datos, usando x como la variable independiente. (*c*) Encuentre la recta de mínimos cuadrados que se ajuste a los datos, usando y como la variable independiente. (*d*) Si un estudiante recibe nota de 75 en álgebra, ¿cuál es su nota esperada en física? (*e*) Si un estudiante recibe nota de 95 en física, ¿cuál es su nota esperada en álgebra?

Tabla 8-32

Álgebra (x)	75	80	93	65	87	71	98	68	84	77
Física (y)	82	78	86	72	91	80	95	72	89	74

8.67. Refiérase a la tabla 8-33. (*a*) Elabore un diagrama de dispersión. (*b*) Encuentre la recta de regresión de mínimo cuadrados de y en x. (*c*) Encuentre la recta de regresión de mínimos cuadrados de x en y. (*d*) Elabore una gráfic de las dos rectas de regresión de (*b*) y (*c*) sobre el diagrama de dispersión de (*a*).

Tabla 8-33

Nota en la primera evaluación (x)	6	5	8	8	7	6	10	4	9	7
Nota en la segunda evaluación (y)	8	7	7	10	5	8	10	6	8	6

CURVAS DE REGRESIÓN DE MÍNIMOS CUADRADOS

8.68. Ajuste una parábola de mínimos cuadrados, $y = a + bx + cx^2$ a los datos de la tabla 8-34.

Tabla 8-34

x	0	1	2	3	4	5	6
y	2.4	2.1	3.2	5.6	9.3	14.6	21.9

8.69. La tabla 8-35 da la distancia d (pies) a la que un automóvil que viaja a velocidad v (millas por hora) se detiene, una vez el conductor detecta un peligro. (*a*) Dibuje d contra v. (*b*) Ajuste una parábola de mínimos cuadrados a la forma $d = a + bv + cv^2$ a los datos. (*c*) Estime d cuando $v = 45$ millas por hora y 80 millas por hora.

Tabla 8-35

Velocidad, v (millas por hora)	20	30	40	50	60	70
Distancia a la que se detiene, d (pies)	54	90	138	206	292	396

8.70. En la tabla 8-36 se da el número y de bacterias por unidad de volumen presente en un cultivo después de x horas. (*a*) Dibuje los datos en papel semilogarítmico, usando la escala logarítmica para y y la escala aritmética para x. (*b*) Ajuste una curva de mínimos cuadrados con la forma $y = ab^x$ a los datos, y explique por qué esta ecuación en particular puede dar buenos resultados. (*c*) Compare con los valores reales los valores de y obtenidos a partir de esta ecuación. (*d*) Estime el valor de y cuando $x = 7$.

Tabla 8-36

Número de horas (x)	0	1	2	3	4	5	6
Número de bacterias por unidad de volumen (y)	32	47	65	92	132	190	275

REGRESIÓN MÚLTIPLE

8.71. La tabla 8-37 muestra los valores correspondientes a tres variables x, y y z. (*a*) Encuentre la ecuación de regresión lineal de mínimos cuadrados de z en x y en y. (*b*) Estime z cuando $x = 10$ y $y = 6$.

Tabla 8-37

x	3	5	6	8	12	14
y	16	10	7	4	3	2
z	90	72	54	42	30	12

ERROR ESTÁNDAR DEL ESTIMADO Y COEFICIENTE DE CORRELACIÓN LINEAL

8.72. Encuentre (*a*) $S_{y.x}$, (*b*) $S_{x.y}$ para los datos del problema 8.67.

8.73. Calcule (*a*) la variación total de y, (*b*) la variación no explicada de y, (*c*) la variación explicada de y para los datos del problema 8.67.

8.74. Use los resultados del problema 8.73 para encontrar el coeficiente de correlación entre los dos conjuntos de notas de las evaluaciones del problema 8.67.

8.75. Encuentre la covarianza de los datos del problema 8.67 (*a*) directamente, (*b*) usando la fórmula $s_{xy} = rs_x s_y$ y el resultado del problema 8.74.

8.76. La tabla 8-38 muestra las edades x y la presión sistólica y de 12 mujeres. (*a*) Encuentre el coeficiente de correlación entre x y y. (*b*) Determine la recta de regresión de mínimos cuadrados y en x. (*c*) Calcule la presión sanguínea de una mujer cuya edad es 45 años.

Tabla 8-38

Edad (x)	56	42	72	36	63	47	55	49	38	42	68	60
Presión sanguínea (y)	147	125	160	118	149	128	150	145	115	140	152	155

8.77. Encuentre los coeficientes de correlación de los datos (*a*) del problema 8.64, (*b*) del problema 8.66.

8.78. El coeficiente de correlación entre dos variables x y y es $r = 0.60$. Si $s_x = 1.50$, $s_y = 2.00$, $\bar{x} = 10$ y $\bar{y} = 20$, encuentre las ecuaciones de las rectas de regresión de (*a*) y en x, (*b*) x en y.

8.79. Calcule (*a*) $s_{y.x}$, (*b*) $s_{x.y}$ para los datos del problema 8.78.

8.80. Si $s_{y \cdot x} = 3$ y $s_y = 5$, encuentre r.

8.81. Si el coeficiente de correlación entre x y y es 0.50, ¿qué porcentaje de la variación total permanece sin explicación por la ecuación de regresión?

8.82. (*a*) Calcule el coeficiente de correlación entre los valores correspondientes de x y y dados en la tabla 8-39. (*b*) Multiplique cada valor de x en la tabla por 2 y sume 6. Multiplique cada valor de y en la tabla por 3 y reste 15. Encuentre el coeficiente de correlación entre los dos nuevos conjuntos de valores, explicando por qué usted obtiene o no el mismo resultado de la parte (*a*).

Tabla 8-39

x	2	4	5	6	8	11
y	18	12	10	8	7	5

COEFICIENTE DE CORRELACIÓN GENERALIZADO

8.83. Encuentre el error estándar del estimado de z en y para los datos del problema 8.71.

8.84. Calcule el coeficiente de correlación múltiple para los datos del problema 8.71.

CORRELACIÓN DE RANGO

8.85. En un torneo se pide a dos jueces que clasifiquen 8 candidatos *A, B, C, D, E, F, G* y *H* según su orden de preferencia. En la tabla 8-40 se presentan sus elecciones. Encuentre el coeficiente de correlación de rango y decida qué tanto concuerdan los jueces en sus juicios.

Tabla 8-40

Candidato	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
Primer juez	5	2	8	1	4	6	3	7
Segundo juez	4	5	7	3	2	8	1	6

8.86. Encuentre el coeficiente de correlación de rango para los datos (*a*) del problema 8.67, (*b*) del problema 8.76.

8.87. Encuentre el coeficiente de correlación de rango para los datos del problema 8.62.

TEORÍA MUESTRAL DE LA REGRESIÓN

8.88. Con base en una muestra de tamaño 27 se encontró que la ecuación de regresión de y en x es $y = 25.0 + 2.00x$. Si $s_{y.x} = 1.50$, $s_x = 3.00$ y $\bar{x} = 7.50$, encuentre los límites de confianza del (*a*) 95%, (*b*) 99% para el coeficiente de regresión.

8.89. En el problema 8.88, pruebe la hipótesis de que el coeficiente de regresión de la población es (*a*) tan pequeño como 1.70, (*b*) tan alto como 2.20 al nivel de significancia del 0.01.

8.90. En el problema 8.88 encuentre los límites de confianza del (*a*) 95%, (*b*) 99% para y cuando $x = 6.00$.

8.91. En el problema 8.88 encuentre los límites de confianza del (*a*) 95%, (*b*) 99% para la media de todos los valores de y correspondientes a $x = 6.00$.

8.92. En relación con el problema 8.76, encuentre los límites de confianza del 95% para (*a*) el coeficiente de regresión de y en x, (*b*) la presión sanguínea de todas las mujeres cuya edad es 45 años, (*c*) la media de la presión sanguínea de todas las mujeres de 45 años.

TEORÍA MUESTRAL DE LA CORRELACIÓN

8.93. Se calculó que un coeficiente de correlación basado en una muestra de tamaño 27 fue 0.40. ¿Podemos concluir al nivel de significancia de (*a*) 0.05, (*b*) 0.01 que el correspondiente coeficiente de correlación de la población es significativamente mayor que cero?

8.94. Se calculó que un coeficiente de correlación basado en una muestra de tamaño 35 fue 0.50. ¿Podemos rechazar la hipótesis de que el coeficiente de correlación de la población es (*a*) tan pequeño como $\rho = 0.30$, (*b*) tan grande como $\rho = 0.70$, usando un nivel de significancia de 0.05?

8.95. Encuentre los límites de confianza del (*a*) 95%, (*b*) 99% para un coeficiente de correlación, el cual se calculó en 0.60 a partir de una muestra de tamaño 28.

8.96. Desarrolle el problema 8.95 si el tamaño de la muestra es 12.

8.97. Encuentre los límites de confianza del 95% para el coeficiente de correlación calculado en el problema 8.76.

8.98. Se calculó que dos coeficientes de correlación a partir de muestras de tamaño 23 y 28, fueron 0.80 y 0.95, respectivamente. ¿Podemos concluir al nivel de (*a*) 0.05, (*b*) 0.01, que hay una diferencia significativa entre ambos coeficientes?

PROBLEMAS DIVERSOS

8.99. Las rectas de regresión de mínimos cuadrados muestrales para un conjunto de datos que involucran X y Y están dadas por $2x - 5y = 3$, $5x - 8y = 2$. Encuentre el coeficiente de correlación lineal.

8.100. Encuentre el coeficiente de correlación entre las estaturas y los pesos de 300 adultos varones en Estados Unidos, tal como aparece en la tabla 8-41.

Tabla 8-41

PESOS y (libras) \ ESTATURAS x (pulgadas)	59–62	63–66	67–70	71–74	75–78
90–109	2	1			
110–129	7	8	4	2	
130–149	5	15	22	7	1
150–169	2	12	63	19	5
170–189		7	28	32	12
190–209		2	10	20	7
210–229			1	4	2

8.101. (*a*) Encuentre la recta de regresión de mínimos cuadrados de y en x para los datos del problema 8.100. (*b*) Estime los pesos de dos hombres cuyas estaturas son 64 y 72 pulgadas, respectivamente.

8.102. Encuentre (*a*) $s_{y.x}$, (*b*) $s_{x.y}$ para los datos del problema 8.100.

8.103. Encuentre los límites de confianza para el coeficiente de correlación calculado en el problema 8.100.

8.104. Encuentre el coeficiente de correlación entre los índices de precios al consumidor en Estados Unidos y los índices de precios de ventas al por mayor para todos los artículos, como se muestra en la tabla 8-42. El periodo base es $1947-1949 = 100$

Tabla 8-42

Año	1949	1950	1951	1952	1953	1954	1955	1956	1957	1958
Índice de precios al consumidor	101.8	102.8	111.0	113.5	114.4	114.8	114.5	116.2	120.2	123.5
Índice de precios por ventas al por mayor	99.2	103.1	114.8	111.6	110.1	110.3	110.7	114.3	117.6	119.2

Fuente: Bureau of Labor Statistics.

8.105. Refiérase a la tabla 8-43. (*a*) Dibuje los datos. (*b*) Encuentre la recta de mínimos cuadrados que se ajusta a los datos y dibújela. (*c*) Calcule la tendencia de precios y compárelos con los precios reales. (*d*) Prediga el índice de precios para la asistencia médica durante 1958 y compárelo con el valor real (144.4). (*e*) ¿En qué año podemos esperar que el índice de costos médicos sea el doble del de 1947 a 1949, suponiendo que continúen las tendencias actuales?

Tabla 8-43

Año	1950	1951	1952	1953	1954	1955	1956	1957
Índice de precios al consumidor para asistencia médica (1947-1949 = 100)	106.0	111.1	117.2	121.3	125.2	128.0	132.6	138.0

Fuente: Bureau of Labor Statistics.

8.106. Refiérase a la tabla 8-44. (*a*) Dibuje los datos. (*b*) Encuentre la parábola de mínimos cuadrados que se ajusta a los datos. (*c*) Calcule los valores de tendencia y compárelos con los valores reales. (*d*) Explique por qué la ecuación obtenida en (*b*) no es útil con fines de extrapolación.

Tabla 8-44

Año	1915	1920	1925	1930	1935	1940	1945	1950	1955
Tasa de natalidad por cada 1000 habitantes	25.0	23.7	21.3	18.9	16.9	17.9	19.5	23.6	24.6

Fuente: Department of Health and Human Services.

Respuestas a los problemas suplementarios

8.64. (*a*) $y = -\frac{1}{3} + \frac{5}{7}x$ o $y = -0.333 + 0.714x$ (*b*) $x = 1 + \frac{9}{7}y$ o $x = 1.00 + 1.29y$

8.65. (*a*) 3.24, 8.24 (*b*) 10.00 **8.66.** (*b*) $y = 29.13 + 0.661x$ (*c*) $x = -14.39 + 1.15y$ (*d*) 79 (*e*) 95

8.67. (*b*) $y = 4.000 + 0.500x$ (*c*) $x = 2.408 + 0.612y$

8.68. $y = 5.51 + 3.20(x - 3) + 0.733(x - 3)^2$ o $y = 2.51 - 1.20x + 0.733x^2$

8.69. (*b*) $d = 41.77 - 1.096v + 0.08786v^2$ (*c*) 170 pies, 516 pies

8.70. (*b*) $y = 32.14(1.427)^x$ o $y = 32.14(10)^{0.1544x}$ o $y = 32.14e^{0.3556x}$ (*d*) 387

8.71. (*a*) $z = 61.40 - 3.65x + 2.54y$ (*b*) 40

8.72. (*a*) 1.304 (*b*) 1.443

8.73. (*a*) 24.50 (*b*) 17.00 (*c*) 7.50

8.74. 0.5533

8.75. 1.5

8.76. (*a*) 0.8961 (*b*) $y = 80.78 + 1.138x$ (*c*) 132

8.77. (*a*) 0.958 (*b*) 0.872

8.78. (*a*) $y = 0.8x + 12$ (*b*) $x = 0.45y + 1$

8.79. (*a*) 1.60 (*b*) 1.20

8.80. ±0.80

8.81. 75%

8.82. (*a*) −0.9203

8.83. 3.12

8.84. 0.9927

8.85. $r_{\text{rank}} = \frac{2}{3}$

8.86. (*a*) 0.5182 (*b*) 0.9318

8.87. −1.0000

8.88. (*a*) 2.00 ± 0.21 (*b*) 2.00 ± 0.28

8.89. (*a*) Usando una prueba de una cola, podemos rechazar la hipótesis.
(*b*) Usando una prueba de una cola, no podemos rechazar la hipótesis.

8.90. (*a*) 37.0 ± 3.6 (*b*) 37.0 ± 4.9

8.91. (*a*) 37.0 ± 1.5 (*b*) 37.0 ± 2.1

8.92. (*a*) 1.138 ± 0.398 (*b*) 132.0 ± 19.2 (*c*) 132.0 ± 5.4

8.93. (*a*) Sí. (*b*) No.

8.94. (*a*) No. (*b*) Sí.

8.95. (*a*) 0.2923 y 0.7951
(*b*) 0.1763 y 0.8361

8.96. (*a*) 0.3912 y 0.7500
(*b*) 0.3146 y 0.7861

8.97. 0.7096 y 0.9653

8.98. (*a*) Sí (*b*) No

8.99. 0.8

8.100. 0.5440

8.101. (*a*) $y = 4.44x - 142.22$ (*b*) 141.9 y 177.5 libras

8.102. (*a*) 16.92 lb (*b*) 2.07 pulgadas

8.103. 0.4961 y 0.7235

8.104. 0.9263

8.105. (*b*) $y = 122.42 + 2.19x$ si la unidad x es $\frac{1}{2}$ año y el origen es enero 1 de 1954; o
$y = 107.1 + 4.38x$ si la unidad x es 1 año y el origen es julio 1 de 1950
(*d*) 142.1 (*e*) 1971

8.106. (*b*) $y = 18.16 - 0.1083x + 0.4653x^2$, donde y es la tasa de natalidad de cada 1000 habitantes y la unidad x es 5 años con origen en julio 1 de 1935

CAPÍTULO 9

Análisis de varianza

EL PROPÓSITO DEL ANÁLISIS DE VARIANZA

En el capítulo 7 usamos la teoría del muestreo para probar la significancia de las diferencias entre dos medias muestrales. Supusimos que las dos poblaciones de las que se tomaron las muestras tenían la misma varianza. En muchas situaciones existe la necesidad de probar la significancia de las diferencias entre tres o más medias muestrales, o de manera equivalente, probar la hipótesis nula de que las medias muestrales son iguales.

EJEMPLO 9.1. Supongamos que en un experimento agrícola, cuatro tratamientos químicos del suelo producen, respectivamente, 28, 22, 18 y 24 toneladas por hectárea. ¿Hay una diferencia significativa en estos promedios, o la dispersión observada se debe simplemente al azar?

Los problemas como éstos se pueden resolver usando una técnica importante conocida como el *análisis de varianza*, desarrollado por Fisher. Este análisis hace uso de la distribución F, de la que hablamos en capítulos anteriores.

CLASIFICACIÓN A UNA VÍA O EXPERIMENTOS DE UN FACTOR

En los *experimentos de un factor* las medidas u observaciones se obtienen para a grupos independientes de muestras, donde el número de medidas en cada grupo es b. Hablando de los a tratamientos, cada uno de los cuales tiene b *repeticiones* o *réplicas*. En el ejemplo 9.1, $a = 4$.

Los resultados de un experimento de un factor se pueden presentar en una tabla que tenga a filas y b columnas (tabla 9-1). Aquí x_{jk} denota la medida de la j-ésima fila y de la k-ésima columna, donde $j = 1, 2, \ldots, a$ y $k = 1, 2, \ldots, b$. Por ejemplo, x_{35} se refiere a la quinta medición del tercer tratamiento.

Tabla 9-1

Tratamiento 1	x_{11}	x_{12}	$\cdots$	x_{1b}	$\bar{x}_{1.}$
Tratamiento 2	x_{21}	x_{22}	$\cdots$	x_{2b}	$\bar{x}_{2.}$
$\vdots$			$\vdots$		
Tratamiento a	x_{a1}	x_{a2}	$\cdots$	x_{ab}	$\bar{x}_{a.}$

Podemos denotar por $\bar{x}_{j.}$ la media de las medidas en la j-ésima fila. Tenemos

$$\bar{x}_{j.} = \frac{1}{b}\sum_{k=1}^{b} x_{jk} \; ; \qquad j = 1, 2, \ldots, a \tag{1}$$

El punto en $\bar{x}_{j.}$ se utiliza para mostrar que se sumó sobre el índice k. Los valores $\bar{x}_{j.}$ se llaman *medias de grupo*, *medias de tratamientos* o *medias de fila*. La *gran media*, o *media total*, es la media de todas las medidas en todos los grupos y se denota por $\bar{x}$, a saber,

$$\bar{x} = \frac{1}{ab}\sum_{j,k} x_{jk} = \frac{1}{ab}\sum_{j=1}^{a}\sum_{k=1}^{b} x_{jk} \tag{2}$$

VARIACIÓN TOTAL. VARIACIÓN DENTRO DE TRATAMIENTOS. VARIACIÓN ENTRE TRATAMIENTOS

Definimos la *variación total*, denotada por v, como la suma de los cuadrados de las desviaciones de cada medida a partir de la gran media $\bar{x}$, es decir,

$$\text{Variación total} = v = \sum_{j,k} (x_{jk} - \bar{x})^2 \tag{3}$$

Escribiendo la identidad,

$$x_{jk} - \bar{x} = (x_{jk} - \bar{x}_{j.}) + (\bar{x}_{j.} - \bar{x}) \tag{4}$$

y luego elevando al cuadrado y sumando sobre j y k, podemos demostrar (*véase* el problema 9.1) que

$$\sum_{j,k} (x_{jk} - \bar{x})^2 = \sum_{j,k} (x_{jk} - \bar{x}_{j.})^2 + \sum_{j,k} (\bar{x}_{j.} - \bar{x})^2 \tag{5}$$

o

$$\sum_{j,k} (x_{jk} - \bar{x})^2 = \sum_{j,k} (x_{jk} - \bar{x}_{j.})^2 + b\sum_{j} (\bar{x}_{j.} - \bar{x})^2$$

Llamamos la primera sumatoria a la derecha de (5) o (6) *variación dentro de tratamientos* (ya que involucra a los cuadrados de las desviaciones de x_{jk} con respecto a las medias de tratamiento $\bar{x}_{j.}$) y lo denotamos por ν_w. Por consiguiente,

$$v_w = \sum_{j,k} (x_{jk} - \bar{x}_{j.})^2 \tag{7}$$

la segunda sumatoria a la derecha de (5) o (6) se llama *variación entre tratamientos* (ya que involucra los cuadrados de las desviaciones de las diferentes medias de tratamiento $\bar{x}_{j.}$ con respecto a la gran media $\bar{x}$) y se denota por ν_b. Por consiguiente,

$$v_b = \sum_{j,k} (\bar{x}_{j.} - \bar{x})^2 = b\sum_{j} (\bar{x}_{j.} - \bar{x})^2 \tag{8}$$

Por lo tanto, las ecuaciones (5) y (6) se pueden escribir

$$v = v_w + v_b \tag{9}$$

MÉTODOS ABREVIADOS PARA OBTENER VARIACIONES

Para reducir al máximo el trabajo de calcular las variaciones anteriores, conviene utilizar:

$$v = \sum_{j,k} x_{jk}^2 - \frac{\tau^2}{ab} \tag{10}$$

$$v_b = \frac{1}{b}\sum_j \tau_{j.}^2 - \frac{\tau^2}{ab} \tag{11}$$

$$v_w = v - v_b \tag{12}$$

donde τ es el total de todos los valores x_{jk} y τ_j es el total de todos los valores del j-ésimo tratamiento, es decir,

$$\tau = \sum_{j,k} x_{jk} \qquad \tau_{j.} = \sum_k x_{jk} \tag{13}$$

En la práctica es conveniente restar algunos valores fijos a todos los datos de la tabla; esto no afecta los resultados finales.

MODELO MATEMÁTICO LINEAL PARA EL ANÁLISIS DE VARIANZA

Podemos considerar que cada fila de la tabla 9-1 representa una muestra aleatoria de tamaño b de la población para ese tratamiento en particular. Por consiguiente, para cada tratamiento j tenemos las variables aleatorias independientes e idénticamente distribuidas $X_{j1}, X_{j2}, \ldots X_{jb}$, las cuales, respectivamente, toman los valores $x_{j1}, x_{j2}, \ldots, x_{jb}$. Cada una de las X_{jk} $(k = 1, 2, \ldots, b)$ se puede expresar como la suma de su valor esperado y un término de "posibilidad" o "error":

$$X_{jk} = \mu_j + \Delta_{jk} \tag{14}$$

Las Δ_{jk} se pueden tomar como variables aleatorias independientes (con relación a j y a k) con distribución normal, con media cero y varianza σ^2. Esto equivale a suponer que las variables X_{jk} $(j = 1, 2, \ldots, a; k = 1, 2, \ldots, b)$ son mutuamente independientes y normales con medias μ_j y varianza común σ^2.

Definamos la constante μ por

$$\mu = \frac{1}{a}\sum_j \mu_j$$

Podemos pensar en μ como la media para un tipo de gran población que comprende todas las poblaciones de tratamientos. Entonces (*14*) se puede reescribir como (*véase* el problema 9.18)

$$X_{jk} = \mu + \alpha_j + \Delta_{jk} \quad \text{donde} \quad \sum_j \alpha_j = 0 \tag{15}$$

La constante α_j puede considerarse como el efecto especial de j-ésimo tratamiento.

La hipótesis nula de que todas las medias de los tratamientos son iguales está dada por $(H_0: \alpha_j = 0; j = 1, 2, \ldots, a)$ o de manera equivalente por $(H_0: \mu_j = \mu; j = 1, 2, \ldots, a)$. Si H_0 es verdadera, las poblaciones de tratamiento, que suponemos normales, tienen media común y varianza común. Entonces hay sólo una población de tratamiento, y todos los tratamientos son estadísticamente idénticos.

VALORES ESPERADOS DE LAS VARIACIONES

La variación entre tratamientos V_b, la variación dentro de tratamientos V_w y la variación total V son variables aleatorias que asumen, respectivamente, los valores ν_b, ν_w y ν, tal como se definió en (*8*), (*7*) y (*3*). Podemos demostrar (problema 9.19) que

$$E(V_b) = (a-1)\sigma^2 + b\sum_j \alpha_j^2 \tag{16}$$

$$E(V_w) = a(b-1)\sigma^2 \tag{17}$$

$$E(V) = (ab-1)\sigma^2 + b\sum \alpha_j^2 \tag{18}$$

A partir de (*17*) se deduce que

$$E\left[\frac{V_w}{a(b-1)}\right] = \sigma^2 \tag{19}$$

de manera que

$$\hat{S}_w^2 = \frac{V_w}{a(b-1)} \tag{20}$$

es siempre el mejor estimador (insesgado) de σ^2 sin importar si H_0 es verdadera o no. Por otro lado, de (*16*) y (*18*) vemos que sólo si H_0 es verdadera, tendremos

$$E\left(\frac{V_b}{a-1}\right) = \sigma^2 \; ; \qquad E\left(\frac{V}{ab-1}\right) = \sigma^2 \tag{21}$$

de manera que sólo en tal caso

$$\hat{S}_b^2 = \frac{V_b}{a-1} \; ; \qquad \hat{S}^2 = \frac{V}{ab-1} \tag{22}$$

siempre y cuando haya estimadores insesgados de σ^2. Sin embargo, si H_0 no es verdadera, tendremos de (*16*)

$$E(\hat{S}_b^2) = \sigma^2 + \frac{b}{a-1}\sum_j \alpha_j^2 \tag{23}$$

DISTRIBUCIONES DE LAS VARIACIONES

Usando el teorema 4-4, página 125, podemos demostrar los siguientes teoremas fundamentales concernientes a las distribuciones de las variaciones V_w, V_b y V.

Teorema 9-1: V_w/σ^2 tiene distribución chi cuadrado con $a(b-1)$ grados de libertad.

Teorema 9-2: Bajo la hipótesis nula H_0, V_b/σ^2 y V/σ^2 tienen distribución chi cuadrado con $a-1$ y $ab-1$ grados de libertad, respectivamente.

Es importante hacer énfasis en que el teorema 9-1 es válido cumpla o no H_0, mientras que el teorema 9-2 sólo es válido si cumple H_0.

LA PRUEBA *F* PARA LA HIPÓTESIS NULA DE MEDIAS IGUALES

Si la hipótesis nula H_0 no es verdadera, es decir, si las medias de tratamiento no son iguales, vemos por (*23*) que podemos esperar que $\hat{S}_b^2$ sea mayor que σ^2, siendo el efecto más pronunciado a medida que aumenta la discrepancia entre las medias. Por otro lado, de (*19*) y (*20*) podemos esperar que $\hat{S}_w^2$ sea igual a σ^2 sin importar si las medias son iguales o no. Se deduce que $\hat{S}_b^2/\hat{S}_w^2$. Es un buen estadístico para probar la hipótesis nula H_0. Si éste es significativamente grande podemos concluir que hay una diferencia significativa entre las

medias de tratamiento y así rechazar H_0. De otra manera podemos aceptar H_0 o reservarnos el juicio dependiendo de análisis posteriores.

Para usar este estadístico debemos conocer su distribución. Esto se da en el siguiente teorema, el cual es una consecuencia del teorema 5-8, página 172.

Teorema 9-3: El estadístico $F = \hat{S}_b^2/\hat{S}_w^2$ tiene distribución F con $a - 1$ y $a(b - 1)$ grados de libertad.

El teorema 9-3 nos permite probar la hipótesis nula a ciertos niveles específicos de significancia usando una prueba a una cola de la distribución F.

TABLA DE ANÁLISIS DE VARIANZA

Los cálculos requeridos para la prueba anterior se resumen en la tabla 9-2, la cual se denomina *tabla de análisis de varianza*. En la práctica calculamos ν y ν_b usando bien sea el método largo, (*3*) y (*8*), o el método corto, (*10*) y (*11*), y luego calculamos $\nu_w = \nu - \nu_b$. Se debe notar que los grados de libertad para la variación total, es decir, $ab - 1$, es igual a la suma de los grados de libertad para los tratamientos entre y dentro de tratamientos.

Tabla 9-2

Variación	Grados de libertad	Cuadrado medio	F
Entre tratamientos, $\nu_b = b\sum_j (\bar{x}_{j.} - \bar{x})^2$	$a - 1$	$\hat{s}_b^2 = \dfrac{\nu_b}{a-1}$	$\dfrac{\hat{s}_b^2}{\hat{s}_w^2}$ con $a - 1, a(b-1)$ grados de libertad
Dentro de tratamientos, $\nu_w = \nu - \nu_b$	$a(b-1)$	$\hat{s}_w^2 = \dfrac{\nu_w}{a(b-1)}$	
Total, $\nu = \nu_b + \nu_w = \sum_{j,k} (x_{jk} - \bar{x})^2$	$ab - 1$		

MODIFICACIONES PARA NÚMEROS DESIGUALES DE OBSERVACIONES

Los resultados anteriores se pueden modificar con facilidad en caso de que los tratamientos 1, . . ., a tengan números diferentes de observaciones iguales a n_1, . . ., n_a, respectivamente. Por lo tanto obtenemos

$$\nu = \sum_{j,k} (x_{jk} - \bar{x})^2 = \sum_{j,k} x_{jk}^2 - \frac{\tau^2}{n} \tag{24}$$

$$\nu_b = \sum_{j,k} (\bar{x}_{j.} - \bar{x})^2 = \sum_j n_j (\bar{x}_{j.} - \bar{x})^2 = \sum_j \frac{\tau_{j.}^2}{n_j} - \frac{\tau^2}{n} \tag{25}$$

$$\nu_w = \nu - \nu_b \tag{26}$$

donde $\sum_{j,k}$ es la sumatoria sobre k desde 1 hasta n_j, y luego sobre j desde 1 hasta a, $n = \sum_j n_j$ es el número total de observaciones de todos los tratamientos, τ es la suma de todas las observaciones, τ_j es la suma de todos los valores en el j-ésimo tratamiento, y $\sum_j$ es la suma desde $j = 1$ hasta a.

La tabla de análisis de varianza para este caso está dada en la tabla 9-3.

Tabla 9-3

Variación	Grados de libertad	Cuadrado medio	F
Entre tratamientos, $v_b = \sum_j n_j(\bar{x}_{j.} - \bar{x})^2$	$a-1$	$\hat{s}_b^2 = \dfrac{v_b}{a-1}$	$\dfrac{\hat{s}_b^2}{\hat{s}_w^2}$ con $a-1, n-a$ grados de libertad
Dentro de tratamientos, $v_w = v - v_b$	$n-a$	$\hat{s}_w^2 = \dfrac{v_w}{n-a}$	
Total, $v = v_b + v_w = \sum_{j,k}(x_{jk} - \bar{x})^2$	$n-1$		

CLASIFICACIÓN A DOS VÍAS O EXPERIMENTOS DE DOS FACTORES

El concepto del análisis de varianza para la clasificación a una vía o experimentos de un factor se puede generalizar. Ilustramos el procedimiento para la *clasificación a dos vías* o *experimentos de dos factores*.

EJEMPLO 9.2. Supongamos que un experimento agrícola consiste en examinar la producción por hectárea de 4 variedades diferentes de trigo, en donde cada variedad crece en 5 parcelas diferentes. Entonces, se necesita un total de (4)(5) = 20 parcelas. En tal caso es conveniente combinar las parcelas en *bloques*, por ejemplo, 4 parcelas en un bloque, con una variedad diferente de trigo creciendo en cada parcela dentro del bloque. Por consiguiente, aquí se necesitan 5 bloques.

En este caso hay dos clasificaciones o factores, ya que puede haber diferencias en la producción por hectárea debido a (*i*) el tipo particular de trigo o (*ii*) el bloque particular utilizado (el cual puede involucrar diferente fertilidad del suelo, etc.).

Por analogía con el experimento agrícola del ejemplo 9.2, nos referimos a las dos clasificaciones o factores en un experimento como *tratamientos* y *bloques*, pero, por supuesto, nos podemos referir a ellos como factor 1 y factor 2, etc.

NOTACIÓN PARA EXPERIMENTOS DE DOS FACTORES

Suponiendo que tenemos a tratamientos y b bloques, construimos la tabla 9-4, donde se supone que hay un valor experimental (por ejemplo, producción por hectárea) correspondiente a cada tratamiento y bloque. Para un tratamiento j y un bloque k, denotamos este valor como x_{jk}. La media de las entradas en la j-ésima fila se denota como $\bar{x}_j$, donde $j = 1, \ldots, a$, mientras que la media de las entradas en la k-ésima columna se denota por $\bar{x}_k$, donde $k = 1, \ldots, b$. La *media total* o *gran media* se denota como $\bar{x}$. En símbolos,

$$\bar{x}_{j.} = \frac{1}{b}\sum_{k=1}^{b} x_{jk}, \qquad \bar{x}_{.k} = \frac{1}{a}\sum_{j=1}^{a} x_{jk}, \qquad \bar{x} = \frac{1}{ab}\sum_{j,k} x_{jk} \tag{27}$$

Tabla 9-4

		Bloques				
		1	2	...	b	
Tratamientos	1	x_{11}	x_{12}	$\cdots$	x_{1b}	$\bar{x}_{1.}$
	2	x_{21}	x_{22}	$\cdots$	x_{2b}	$\bar{x}_{2.}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots\vdots\vdots$		
	a	x_{a1}	x_{a2}	$\cdots$	x_{ab}	$\bar{x}_{a.}$
		$\bar{x}_{.1}$	$\bar{x}_{.2}$	$\cdots$	$\bar{x}_{.b}$	

VARIACIONES PARA EXPERIMENTOS DE DOS FACTORES

Tal como en el caso de los experimentos de un factor, podemos definir las variaciones para los experimentos de dos factores. Primero definimos la *variación total*, como en (*3*), por

$$v = \sum_{j,k} (x_{jk} - \bar{x})^2 \tag{28}$$

Al escribir la identidad

$$x_{jk} - \bar{x} = (x_{jk} - \bar{x}_{j.} - \bar{x}_{.k} + \bar{x}) + (\bar{x}_{j.} - \bar{x}) + (\bar{x}_{.k} - \bar{x}) \tag{29}$$

y luego elevando al cuadrado y sumando sobre j y k, podemos demostrar que

$$v = v_e + v_r + v_c \tag{30}$$

donde ν_e = variación debida al error o al azar $= \sum_{j,k} (x_{jk} - \bar{x}_{j.} - \bar{x}_{.k} + \bar{x})^2$

ν_r = variación entre filas (tratamientos) $= b \sum_{j=1}^{a} (\bar{x}_{j.} - \bar{x})^2$

ν_c = variación entre columnas (bloques) $= a \sum_{k=1}^{b} (\bar{x}_{.k} - \bar{x})^2$

La variación debida al error o al azar también se conoce como *variación residual*.

Las siguientes son fórmulas cortas para el cálculo, análogas a (*10*), (*11*) y (*12*)

$$v = \sum_{j,k} x_{jk}^2 - \frac{\tau^2}{ab} \tag{31}$$

$$v_r = \frac{1}{b} \sum_{j=1}^{a} \tau_{j.}^2 - \frac{\tau^2}{ab} \tag{32}$$

$$v_c = \frac{1}{a} \sum_{k=1}^{b} \tau_{.k}^2 - \frac{\tau^2}{ab} \tag{33}$$

$$v_e = v - v_r - v_c \tag{34}$$

donde $\tau_{j.}$ es el total de entradas de la j-ésima fila, $\tau_{.k}$ es el total de entradas de la k-ésima columna y τ es el total de todas las entradas.

ANÁLISIS DE VARIANZA PARA EXPERIMENTOS DE DOS FACTORES

Para el modelo matemático de experimentos de dos factores supongamos que las variables aleatorias X_{jk}, cuyos valores son x_{jk} se pueden escribir como

$$X_{jk} = \mu + \alpha_j + \beta_k + \Delta_{jk} \tag{35}$$

Aquí μ es la gran media de la población, α_j es la parte de X_{jk} debida a la diferencia de tratamientos (algunas veces llamada *efecto de tratamientos*), β_k es la parte debida a la diferencia de bloques (algunas veces llamada *efecto de bloques*) y Δ_{jk} es la parte de X_{jk} debida al azar o al error. Como lo hicimos antes, podemos tomar las Δ_{jk} como variables aleatorias independientes con distribución normal con media cero y desviación estándar σ^2, de manera que los X_{jk} también son variables independientes con distribución normal con varianza σ^2. Bajo suposiciones adecuadas sobre las medias de X_{jk} tenemos

$$\sum_j \alpha_j = 0 \qquad \sum_k \beta_k = 0 \tag{36}$$

que hace

$$\mu = \frac{1}{ab}\sum_{j,k} E(X_{jk})$$

Con relación a los resultados de (*16*) a (*18*), podemos demostrar que

$$E(V_r) = (a-1)\sigma^2 + b\sum_j \alpha_j^2$$

$$E(V_c) = (b-1)\sigma^2 + a\sum_k \beta_k^2 \tag{37}$$

$$E(V_e) = (a-1)(b-1)\sigma^2 \tag{38}$$

$$E(V) = (ab-1)\sigma^2 + b\sum_j \alpha_j^2 + a\sum_k \beta_k^2 \tag{39}$$

Hay dos hipótesis nulas que queremos probar: (*40*)

$H_0^{(1)}$: Todas las medias de tratamientos (filas) son iguales, es decir, $\alpha_j = 0$, $j = 1, \ldots, a$

$H_0^{(2)}$: Todas las medias de bloques (columnas) son iguales, es decir, $\beta_k = 0$, $k = 1, \ldots, b$

De (*39*) vemos que, sin tener en cuenta $H_0^{(1)}$ o $H_0^{(2)}$, un mejor estimador (insesgado) de σ^2 está dado por

$$\hat{S}_e^2 = \frac{V_e}{(a-1)(b-1)} \quad \text{es decir,} \quad E(\hat{S}_e^2) = \sigma^2 \tag{41}$$

También, si las hipótesis $H_0^{(1)}$ y $H_0^{(2)}$ son verdaderas, entonces

$$\hat{S}_r^2 = \frac{V_r}{a-1}, \qquad \hat{S}_c^2 = \frac{V_c}{b-1}, \qquad \hat{S}^2 = \frac{V}{ab-1} \tag{42}$$

Serán estimadores insesgados de σ^2. Si $H_0^{(1)}$ y $H_0^{(2)}$ no son verdaderas, sin embargo, tenemos de (*37*) y (*38*), respectivamente,

$$E(\hat{S}_r^2) = \sigma^2 + \frac{b}{a-1}\sum_j \alpha_j^2 \tag{43}$$

$$E(\hat{S}_c^2) = \sigma^2 + \frac{a}{b-1}\sum_k \beta_k^2$$

Los siguientes teoremas son similares a los teoremas 9-1 y 9-2. (*44*)

Teorema 9-4: V_e/σ^2 tiene distribución chi cuadrado con $(a-1)(b-1)$ grados de libertad, sin tener en cuenta $H_0^{(1)}$ o $H_0^{(2)}$.

Teorema 9-5: Bajo la hipótesis $H_0^{(1)}$, V_r/σ^2 tiene distribución chi cuadrado con $a-1$ grados de libertad. Bajo la hipótesis $H_0^{(2)}$, V_c/σ^2 tiene distribución chi cuadrado con $b-1$ grados de libertad. Bajo ambas hipótesis $H_0^{(1)}$ y $H_0^{(2)}$, V/σ^2 tiene distribución chi cuadrado con $ab-1$ grados de libertad.

Para probar la hipótesis $H_0^{(1)}$ es natural considerar el estadístico $\hat{S}_r^2/\hat{S}_e^2$ ya que de (*43*) vemos que se espera que $\hat{S}_r^2$ difiera significativamente de σ^2 si las medias de la fila (tratamiento) son significativamente diferentes.

De igual manera, para probar la hipótesis $H_0^{(2)}$, consideramos el estadístico $\hat{S}_c^2/\hat{S}_e^2$. Las distribuciones de $\hat{S}_r^2/\hat{S}_e^2$ y de $\hat{S}_c^2/\hat{S}_e^2$ están dadas en la siguiente analogía del teorema 9-3.

Teorema 9-6: Bajo la hipótesis $H_0^{(1)}$ el estadístico $\hat{S}_r^2/\hat{S}_e^2$ tiene distribución F con $a-1$ y $(a-1)(b-1)$ grados de libertad. Bajo la hipótesis $H_0^{(2)}$ el estadístico $\hat{S}_c^2/\hat{S}_e^2$ tiene distribución F con $b-1$ y $(a-1)(b-1)$ grados de libertad.

El teorema nos permite aceptar o rechazar $H_0^{(1)}$ y $H_0^{(2)}$ a niveles especificados de significancia. Por conveniencia, como en el caso de un factor, se puede elaborar una tabla de análisis de varianza, como se muestra en la tabla 9-5.

EXPERIMENTOS DE DOS FACTORES CON REPLICACIÓN

En la tabla 9-4 hay sólo una entrada correspondiente a un tratamiento dado y a un bloque dado. Con frecuencia, repitiendo el experimento se puede obtener más información relacionada con los factores, un proceso llamado *replicación*. En tal caso habrá más de una entrada correspondiente a un tratamiento dado y a un bloque dado. Debemos suponer que hay c entradas para cada posición; se pueden efectuar los cambios apropiados cuando las cantidades de replicaciones no son iguales.

Debido a la replicación, es preciso usar un modelo adecuado que remplace el modelo dado en (*35*), página 341. Para obtenerlo, decimos que X_{jkl} denota la variable aleatoria correspondiente a la j-ésima fila o tratamiento, la k-ésima columna o bloque y la l-ésima repetición o replicación. El modelo está dado por

$$X_{jkl} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \Delta_{jkl}$$

donde μ, α_j, β_k se definieron anteriormente, Δ_{jkl} son las variables aleatorias independientes con distribución normal, media cero y varianza σ^2, mientras que γ_{jk} denota los *efectos de interacción* (o simplemente *interacciones*) fila-columna o tratamiento-bloque. Correspondiendo a (*36*) tenemos

$$\sum_j \alpha_j = 0, \qquad \sum_k \beta_k = 0, \qquad \sum_j \gamma_{jk} = 0, \qquad \sum_k \gamma_{jk} = 0 \tag{45}$$

Tabla 9-5

Variación	Grados de libertad	Cuadrado medio	F
Entre tratamientos, $v_r = b\sum_j (\bar{x}_{j.} - \bar{x})^2$	$a-1$	$\hat{s}_r^2 = \dfrac{v_r}{a-1}$	$\hat{s}_r^2/\hat{s}_e^2$ con $a-1$, $(a-1)(b-1)$ grados de libertad
Entre bloques, $v_c = a\sum_k (\bar{x}_{.k} - \bar{x})^2$	$b-1$	$\hat{s}_c^2 = \dfrac{v_r}{b-1}$	$\hat{s}_c^2/\hat{s}_e^2$ con $b-1$, $(a-1)(b-1)$ grados de libertad
Residuo o azar, $v_e = v - v_r - v_c$	$(a-1)(b-1)$	$\hat{s}_e^2 = \dfrac{v_e}{(a-1)(b-1)}$	
Total, $v = v_r + v_c + v_e = \sum_{j,k} (x_{jk} - \bar{x})^2$	$ab-1$		

Igual que antes, la variación total v de todos los datos se puede dividir en variaciones debidas a las filas ν_r, a las columnas ν_c, a las replicaciones ν_i y al error aleatorio o residual ν_e:

$$v = v_r + v_c + v_i + v_e \tag{46}$$

$$\text{donde } v = \sum_{j,k,l} (x_{jkl} - \bar{x})^2 \tag{47}$$

$$v_r = bc \sum_{j=1}^{a} (\bar{x}_{j..} - \bar{x})^2 \tag{48}$$

$$v_c = ac \sum_{k=1}^{b} (\bar{x}_{.k.} - \bar{x})^2 \tag{49}$$

$$v_i = c \sum_{j,k} (\bar{x}_{jk.} - \bar{x}_{j..} - \bar{x}_{.k.} + \bar{x})^2 \tag{50}$$

$$v_e = \sum_{j,k,l} (x_{jkl} - \bar{x}_{jk.})^2 \tag{51}$$

Los puntos en los subíndices de estos resultados tienen significados análogos a los que dimos anteriormente (página 336). Por ejemplo,

$$\bar{x}_{j..} = \frac{1}{bc} \sum_{k,l} x_{jkl} = \frac{1}{b} \sum_{k} \bar{x}_{jk.} \tag{52}$$

Usando el número apropiado de grados de libertad para cada fuente de variación, podemos elaborar la tabla de análisis de varianza, tabla 9-6.

Tabla 9-6

Variación	Grados de libertad	Cuadrado medio	F
Entre tratamientos, v_r	$a-1$	$\hat{s}_r^2 = \dfrac{v_r}{a-1}$	$\hat{s}_r^2/\hat{s}_e^2$ con $a-1$, $ab(c-1)$ grados de libertad
Entre bloques, v_c	$b-1$	$\hat{s}_c^2 = \dfrac{v_c}{b-1}$	$\hat{s}_c^2/\hat{s}_e^2$ con $b-1$, $ab(c-1)$ grados de libertad
Interacción, v_i	$(a-1)(b-1)$	$\hat{s}_i^2 = \dfrac{v_i}{(a-1)(b-1)}$	$\hat{s}_i^2/\hat{s}_e^2$ con $(a-1)(b-1)$, $ab(c-1)$ grados de libertad
Residuo o azar, v_e	$ab(c-1)$	$\hat{s}_e^2 = \dfrac{v_e}{ab(c-1)}$	
Total, v	$abc-1$		

Las relaciones F de la última columna de la tabla 9-6 se pueden utilizar para probar las hipótesis nulas

$H_0^{(1)}$: Todas las medias de los tratamientos (fila) son iguales, o sea, $\alpha_j = 0$

$H_0^{(2)}$: Todas las medias de los bloques (columna) son iguales, o sea, $\beta_k = 0$

$H_0^{(3)}$: No hay interacción entre tratamientos y bloques, o sea, $\gamma_{jk} = 0$

Desde un punto de vista práctico podemos decidir primero si rechazar o no $H_0^{(3)}$ a un nivel apropiado de significancia usando la relación $\hat{s}_i^2/\hat{s}_e^2$ de la tabla 9-6. Surgen dos casos posibles.

Caso I: *No se puede rechazar* $H_0^{(3)}$: En este caso podemos concluir que las interacciones no son muy grandes. Podemos probar $H_0^{(1)}$ y $H_0^{(2)}$ usando las relaciones F $\hat{s}_r^2/\hat{s}_e^2$ y $\hat{s}_c^2/\hat{s}_e^2$, respectivamente, como se muestra en la tabla 9-6. Algunos estadísticos recomiendan combinar las variaciones en este caso tomando el total $\nu_i + \nu_e$ y dividiéndolo por el número total de grados de libertad correspondientes, $(a-1)(b-1) + ab(c-1)$, y usando este valor para remplazar el denominador $\hat{s}_e^2$ en la prueba F.

Caso II: *Se puede rechazar* $H_0^{(3)}$: En este caso podemos concluir que las interacciones son significativamente grandes. Las diferencias en los factores serán importantes sólo si son grandes comparadas con tales interacciones. Por esta razón, muchos estadísticos recomiendan que $H_0^{(1)}$ y $H_0^{(2)}$ se prueben usando las relaciones F $\hat{s}_r^2/\hat{s}_i^2$ y $\hat{s}_c^2/\hat{s}_i^2$ en lugar de aquellas dadas en la tabla 9-6. Nosotros también podemos usar este procedimiento alterno.

El análisis de varianza con replicación se lleva a cabo más fácilmente si primero se totalizan los valores correspondientes a tratamientos particulares (filas) y bloques (columnas). Esto produce una tabla de dos factores con entradas sencillas, la cual se puede analizar como en la tabla 9-5. El procedimiento se ilustra en el problema 9.13.

DISEÑO EXPERIMENTAL

Las técnicas del análisis de varianza discutidas anteriormente se usan después de que se obtienen los resultados de un experimento. Sin embargo, con el fin de obtener tanta información como sea posible, los detalles del experimento se deben planear con anterioridad cuidadosamente. A esto se lo denomina *diseño del experimento*. A continuación daremos unos ejemplos importantes del diseño experimental.

1. **ALEATORIEDAD COMPLETA**. Supongamos que tenemos un experimento agrícola como en el ejemplo 9.1, página 335. Para diseñar tal experimento, podemos dividir la tierra en $4 \times 4 = 16$ parcelas (se indica en la figura 9-1 como cuadrados, aunque físicamente la forma puede ser cualquiera) y se les asigna cada tratamiento, indicados por A, B, C, D a los cuatro bloques seleccionados completamente al azar. El propósito de la aleatoriedad es eliminar las fuentes de error, tales como la fertilidad del suelo.

D	A	C	C
B	D	B	A
D	C	B	D
A	B	C	A

Aleatoriedad completa

Fig. 9-1

I	C	B	A	D
II	A	B	D	C
III	B	C	D	A
IV	A	D	C	B

Bloques aleatorios

Fig. 9-2

D	B	C	A
B	D	A	C
C	A	D	B
A	C	B	D

Cuadrado latino

Fig. 9-3

B_γ	A_β	D_δ	C_α
A_δ	B_α	C_γ	D_β
D_α	C_δ	B_β	A_γ
C_β	D_γ	A_α	B_δ

Cuadrado greco latino

Fig. 9-4

2. **BLOQUES ALEATORIOS.** Cuando es necesario tener un conjunto completo de tratamientos para cada bloque, tal como en el ejemplo 9.2, los tratamientos A, B, C, D se introducen en orden aleatorio

dentro de cada bloque I, II, III, IV (*véase* figura 9-2), y por esta razón los bloques se denominan *bloques aleatorios*. Este tipo de diseño se utiliza cuando se quiere controlar *una fuente de error o variabilidad*, a saber, la diferencia en los bloques (filas en la figura 9-2).

3. **CUADRADO LATINO.** Para ciertos propósitos es necesario controlar *dos fuentes de error o variabilidad* al mismo tiempo, tales como las diferencias en las filas y las diferencias en las columnas. Por ejemplo, en el experimento del ejemplo 9.1, los errores en las filas y columnas pudieron haberse debido a cambios en la fertilidad del suelo en diferentes partes de la tierra. En tal caso, es deseable que cada tratamiento ocurra una vez en cada fila y una vez en cada columna, como se ve en la figura 9-3. Este arreglo se llama *cuadrado latino* debido a que se utilizan las letras latinas A, B, C, D.

4. **CUADRADO GRECO LATINO.** Si es necesario controlar *tres* fuentes de error o variabilidad, se debe usar el *cuadrado greco latino*, como se indica en la figura 9-4. Este tipo cuadrado consta básicamente de dos cuadrados latinos, uno sobre otro. Se utilizan las letras latinas A, B, C, D para un cuadrado y las letras griegas α, β, γ, δ para el otro. El requisito adicional es que cada letra latina se use una sola vez con cada letra griega. Cuando se cumple esta propiedad se dice que el cuadrado es *ortogonal*.

Problemas resueltos

CLASIFICACIÓN A UNA VÍA O EXPERIMENTOS DE UN FACTOR

9.1. Demuestre que

$$\sum_{j,k}(x_{jk}-\bar{x})^2=\sum_{j,k}(x_{jk}-\bar{x}_{j.})^2+\sum_{j,k}(\bar{x}_{j.}-\bar{x})^2$$

Tenemos $x_{jk}-\bar{x}=(x_{jk}-\bar{x}_{j.})+(\bar{x}_{j.}-\bar{x})$. Entonces elevando al cuadrado y sumando sobre j y k, encontramos

$$\sum_{j,k}(x_{jk}-\bar{x})^2=\sum_{j,k}(x_{jk}-\bar{x}_{j.})^2+\sum_{j,k}(\bar{x}_{j.}-\bar{x})^2+2\sum_{j,k}(x_{jk}-\bar{x}_{j.})(\bar{x}_{j.}-\bar{x})$$

Para demostrar el resultado requerido, debemos mostrar que la última sumatoria es cero. Con el fin de lograrlo, procedemos como sigue.

$$\sum_{j,k}(x_{jk}-\bar{x}_{j.})(\bar{x}_{j.}-\bar{x})=\sum_{j=1}^{a}(\bar{x}_{j.}-\bar{x})\left[\sum_{k=1}^{b}(x_{jk}-\bar{x}_{j.})\right]$$

$$=\sum_{j=1}^{a}(\bar{x}_{j.}-\bar{x})\left[\left(\sum_{k=1}^{b}x_{jk}\right)-b\bar{x}_{j.}\right]=0$$

ya que $\bar{x}_{j.}=\dfrac{1}{b}\sum_{k=1}^{b}x_{jk}$.

9.2. Verifique que $(a)\ \tau=ab\bar{x}$, $(b)\ \tau_{j.}=b\bar{x}_{j.}$, $(c)\ \sum_j \tau_{j.}=ab\bar{x}$ usando la notación de la página 337.

(*a*)
$$\tau=\sum_{j,k}x_{jk}=ab\left(\frac{1}{ab}\sum_{j,k}x_{jk}\right)=ab\bar{x}$$

(*b*)
$$\tau_{j.}=\sum_{k}x_{jk}=b\left(\frac{1}{b}\sum_{k}x_{jk}\right)=b\bar{x}_{j.}$$

(*c*) ya que $\tau_{j.}=\sum_k x_{jk}$, tenemos

$$\sum_j \tau_{j.}=\sum_j\sum_k x_{jk}=\tau=ab\bar{x}$$

de la parte (*a*).

9.3. Verifique las fórmulas abreviadas de (*10*) a (*12*), página 337.

Tenemos

$$v = \sum_{j,k}(x_{jk} - \bar{x})^2 = \sum_{j,k}(x_{jk}^2 - 2\bar{x}\bar{x}_{jk} + \bar{x}^2)$$
$$= \sum_{j,k} x_{jk}^2 - 2\bar{x}\sum_{j,k} x_{jk} + ab\bar{x}^2$$
$$= \sum_{j,k} x_{jk}^2 - 2\bar{x}(ab\bar{x}) + ab\bar{x}^2$$
$$= \sum_{j,k} x_{jk}^2 - ab\bar{x}^2$$
$$= \sum_{j,k} x_{jk}^2 - \frac{\tau^2}{ab}$$

usando el problema 9.2(*a*) en la tercera y última líneas anteriores. De igual manera

$$v_b = \sum_{j,k}(\bar{x}_{j.} - \bar{x})^2 = \sum_{j,k}(\bar{x}_{j.}^2 - 2\bar{x}\bar{x}_{j.} + \bar{x}^2)$$
$$= \sum_{j,k} \bar{x}_{j.}^2 - 2\bar{x}\sum_{j,k} \bar{x}_{j.} + ab\bar{x}^2$$
$$= \sum_{j,k}\left(\frac{\tau_{j.}}{b}\right)^2 - 2\bar{x}\sum_{j,k}\frac{\tau_{j.}}{b} + ab\bar{x}^2$$
$$= \frac{1}{b^2}\sum_{j=1}^{a}\sum_{k=1}^{b}\tau_{j.}^2 - 2\bar{x}(ab\bar{x}) + ab\bar{x}^2$$
$$= \frac{1}{b}\sum_{j=1}^{a}\tau_{j.}^2 - ab\bar{x}^2$$
$$= \frac{1}{b}\sum_{j=1}^{a}\tau_{j.}^2 - \frac{\tau^2}{ab}$$

usando el problema 9.2(*b*) en la tercera línea y el problema 9.2(*a*) en la última línea.

Finalmente, (*12*) se deduce del hecho de que $\nu = \nu_b + \nu_w$ o $\nu_w = \nu - \nu_b$.

9.4. La tabla 9-7 muestra la producción en toneladas por hectárea de cierta variedad de trigo que crece en un tipo particular de suelo tratado con los químicos A, B o C. Encuentre (*a*) la producción media para los diferentes tratamientos, (*b*) la gran media para todos los tratamientos, (*c*) la variación total, (*d*) la variación entre tratamientos, (*e*) la variación dentro de tratamientos. Utilice el método largo.

Tabla 9-7

A	48	49	50	49
B	47	49	48	48
C	49	51	50	50

Tabla 9-8

3	4	5	4
2	4	3	3
4	6	5	5

Para simplificar la aritmética, restemos algunos números apropiados, como 45, a todos los datos sin afectar los valores de las variaciones. Entonces obtenemos los datos de la tabla 9-8.

(*a*) Las medias de los tratamientos (filas) para la tabla 9-8 se dan, respectivamente, por

$$\bar{x}_{1.} = \frac{1}{4}(3+4+5+4) = 4, \qquad \bar{x}_{2.} = \frac{1}{4}(2+4+3+3) = 3, \qquad \bar{x}_{3.} = \frac{1}{4}(4+6+5+5) = 5$$

Por consiguiente, las producciones medias, obtenidas al sumar 45 a éstas, son 49, 48 y 50 toneladas por hectárea para A, B y C, respectivamente.

(*b*) $$\bar{x} = \frac{1}{12}(3+4+5+4+2+4+3+3+4+6+5+5) = 4$$

Por lo tanto, la gran media del conjunto original de datos es 45 + 4 = 49 toneladas por hectárea

(*c*) $$\text{Variación total} = v = \sum_{j,k}(x_{jk} - \bar{x})^2$$

$$= (3-4)^2 + (4-4)^2 + (5-4)^2 + (4-4)^2 + (2-4)^2 + (4-4)^2 + (3-4)^2 + (3-4)^2 + (4-4)^2 + (6-4)^2 + (5-4)^2 + (5-4)^2$$

$$= 14$$

(*d*) $$\text{Variación entre tratamientos} = v_b = b\sum_j(\bar{x}_{j.} - \bar{x})^2$$

$$= 4[(4-4)^2 + (3-4)^2 + (5-4)^2] = 8$$

(*e*) $$\text{Variación dentro de tratamientos} = v_w = v - v_b = 14 - 8 = 6$$

Otro método.

$$v_w = \sum_{j,k}(x_{jk} - \bar{x}_{j.})^2$$

$$= (3-4)^2 + (4-4)^2 + (5-4)^2 + (4-4)^2 + (2-3)^2 + (4-3)^2 + (3-3)^2 + (3-3)^2 + (4-5)^2 + (6-5)^2 + (5-5)^2 + (5-5)^2$$

$$= 6$$

9.5. Con relación al problema 9.4, encuentre un estimador insesgado de la varianza de la población σ^2 a partir de (*a*) la variación entre tratamientos bajo la hipótesis nula de medias de tratamientos iguales, (*b*) la variación dentro de tratamientos.

(*a*) $$\hat{s}_b^2 = \frac{v_b}{a-1} = \frac{8}{3-1} = 4$$

(*b*) $$\hat{s}_w^2 = \frac{v_w}{a(b-1)} = \frac{6}{3(4-1)} = \frac{2}{3}$$

9.6. Con relación al problema 9.4, ¿podemos rechazar la hipótesis nula de medias iguales al nivel de significancia del (*a*) 0.05, (*b*) 0.01?

Tenemos $$F = \frac{\hat{s}_b^2}{\hat{s}_w^2} = \frac{4}{2/3} = 6$$

con $a - 1 = 3 - 1 = 2$ y $a(b-1) = 3(4-1) = 9$ grados de libertad.

(*a*) Refiriéndonos al apéndice F, con $\nu_1 = 2$ y $\nu_2 = 9$, vemos que $F_{0.95} = 4.26$. Dado que $F = 6 > F_{0.95}$, podemos rechazar la hipótesis nula de medias iguales al nivel de 0.05.

(*b*) Refiriéndonos al apéndice F, con $\nu_1 = 2$ y $\nu_2 = 9$, vemos que $F_{0.99} = 8.02$. Dado que $F = 6 < F_{0.99}$, no podemos rechazar la hipótesis nula de medias iguales al nivel de 0.01.

La tabla de análisis de varianza para los problemas de 9.4 a 9.6 se muestra en la tabla 9-9.

Tabla 9-9

Variación	Grados de libertad	Cuadrado medio	F
Entre tratamientos, $v_b = 8$	$a - 1 = 2$	$\hat{s}_b^2 = \frac{8}{2} = 4$	$F = \frac{\hat{s}_b^2}{\hat{s}_w^2} = \frac{4}{2/3} = 6$ con 2,9 grados de libertad
Dentro de tratamientos, $v_w = v - v_b = 14 - 8 = 6$	$a(b-1) = (3)(3) = 9$	$\hat{s}_w^2 = \frac{6}{9} = \frac{2}{3}$	
Total, $v = 14$	$ab - 1 = (3)(4) - 1 = 11$		

9.7. Use las fórmulas abreviadas de (*10*) a (*12*) para obtener los resultados del problema 9.4.

(*a*) Tenemos

$$\sum_{j,k} x_{jk}^2 = 9 + 16 + 25 + 16 + 4 + 16 + 9 + 9 + 16 + 36 + 25 + 25 = 206$$

También

$$\tau = 3 + 4 + 5 + 4 + 2 + 4 + 3 + 3 + 4 + 6 + 5 + 5 = 48$$

Por consiguiente,

$$v = \sum_{j,k} x_{jk}^2 - \frac{\tau^2}{ab} = 206 - \frac{(48)^2}{(3)(4)} = 206 - 192 = 14$$

(*b*) Los totales de las filas son

$$\tau_{1.} = 3 + 4 + 5 + 4 = 16$$
$$\tau_{2.} = 2 + 4 + 3 + 3 = 12$$
$$\tau_{3.} = 4 + 6 + 5 + 5 = 20$$

También

$$\tau = 16 + 12 + 20 = 48$$

Entonces

$$v_b = \frac{1}{b}\sum_j \tau_{j.}^2 - \frac{\tau^2}{ab} = \frac{1}{4}(16^2 + 12^2 + 20^2) - \frac{(48)^2}{(3)(4)} = 200 - 192 = 8$$

(*c*)

$$v_w = v - v_b = 14 - 8 = 6$$

Conviene arreglar los datos como en la tabla 9-10.

Tabla 9-10

					$\tau_{j.}$	$\tau_{j.}^2$
A	3	4	5	4	16	256
B	2	4	3	3	12	144
C	4	6	5	5	20	400
	$\sum_{j,k} x_{jk}^2 = 206$				$\tau = \sum_j \tau_{j.} = 48$	$\sum_j \tau_{j.}^2 = 800$

$$v = 206 - \frac{(48)^2}{(3)(4)} = 14$$

$$v_b = \frac{1}{4}(800) - \frac{(48)^2}{(3)(4)} = 8$$

Los resultados concuerdan con los que se obtuvieron en el problema 9.4, y desde este punto el análisis se hace como anteriormente.

9.8. Una compañía quiere comprar una de cinco máquinas diferentes *A*, *B*, *C*, *D*, *E*. En un experimento diseñado para decidir si existe una diferencia significativa en el desempeño de las máquinas, cinco maquinistas expertos trabajaron con cada una de las máquinas durante tiempos iguales. La tabla 9-11 muestra el número de unidades producidas. Pruebe la hipótesis de que no hay diferencia entre las máquinas al nivel de significancia de (*a*) 0.05, (*b*) 0.01.

Tabla 9-11

A	68	72	75	42	53
B	72	52	63	55	48
C	60	82	65	77	75
D	48	61	57	64	50
E	64	65	70	68	53

Tabla 9-12

						$\tau_{j.}$	$\tau_{j.}^2$
A	8	12	15	−18	−7	10	100
B	12	−8	3	−5	−2	0	0
C	0	22	6	17	15	60	3600
D	−12	1	−3	4	−10	−20	400
E	4	5	10	8	−7	20	400
	$\sum x_{jk}^2 = 2356$					70	4500

Restando a todos los datos un número apropiado, como 60, obtenemos la tabla 9-12.

Entonces
$$v = 2356 - \frac{(70)^2}{(5)(4)} = 2356 - 245 = 2111$$
$$v_b = \frac{1}{5}(4500) - \frac{(70)^2}{(5)(4)} = 900 - 245 = 655$$

Ahora construimos la tabla 9-13.

Tabla 9-13

Variación	Grados de libertad	Cuadrado medio	F
Entre tratamientos, $v_c = 655$	$a - 1 = 4$	$\hat{s}_b^2 = \frac{655}{4} = 163.75$	$F = \frac{\hat{s}_b^2}{\hat{s}_w^2} = 2.25$
Dentro de tratamientos, $v_w = 1456$	$a(b-1) = 5(4) = 20$	$\hat{s}_w^2 = \frac{1456}{(5)(4)} = 72.8$	
Totales, $v = 2111$	$ab - 1 = 24$		

Para 4, 20 grados de libertad tenemos $F_{0.95} = 2.87$. Por consiguiente, no podemos rechazar la hipótesis nula al nivel de 0.05 y por consiguiente, tampoco al nivel de 0.01.

MODIFICACIONES PARA NÚMEROS DESIGUALES DE OBSERVACIONES

9.9. La tabla 9-14 muestra los tiempos de vida en horas de muestras de tres tipos diferentes de tubos de televisión fabricados por una compañía. Usando el método largo, pruebe al nivel de significancia del (*a*) 0.05, (*b*) 0.01 si hay una diferencia entre los tres tipos.

Tabla 9-14

Muestra 1	407	411	409		
Muestra 2	404	406	408	405	402
Muestra 3	410	408	406	408	

Tabla 9-15

						Total	Media
Muestra 1	7	11	9			27	9
Muestra 2	4	6	8	5	2	25	5
Muestra 3	10	8	6	8		32	8
	$\bar{x} =$ gran media $= \frac{84}{12} = 7$						

Es conveniente restar un número apropiado, como 400, y obtenemos la tabla 9-15.
En esta tabla indicamos los totales de las filas, las medias muestrales o de grupo, y la gran media. Entonces tenemos

$$v = \sum_{j,k} (x_{jk} - \bar{x})^2 = (7-7)^2 + (11-7)^2 + \cdots + (8-7)^2 = 72$$

$$v_b = \sum_{j,k} (\bar{x}_{j.} - \bar{x})^2 = \sum_j n_j (\bar{x}_{j.} - \bar{x})^2$$

$$= 3(9-7)^2 + 5(7-5)^2 + 4(8-7)^2 = 36$$

$$v_w = v - v_b = 72 - 36 = 36$$

También obtenemos directamente v_w observando que es igual a

$$(7-9)^2 + (11-9)^2 + (9-9)^2 + (4-5)^2 + (6-5)^2 + (8-5)^2 + (5-5)^2$$
$$+ (2-5)^2 + (10-8)^2 + (8-8)^2 + (6-8)^2 + (8-8)^2$$

Los datos se pueden resumir en la tabla de análisis de varianza, tabla 9-16.

Tabla 9-16

Variación	Grados de libertad	Cuadrado medio	F
$v_b = 36$	$a - 1 = 2$	$\hat{s}_b^2 = \frac{36}{2} = 18$	$\frac{\hat{s}_b^2}{\hat{s}_w^2} = \frac{18}{4} = 4.5$
$v_w = 36$	$n - a = 9$	$\hat{s}_w^2 = \frac{36}{9} = 4$	

Ahora para 2 y 9 grados de libertad encontramos del apéndice F que $F_{0.95} = 4.26$, $F_{0.99} = 8.02$. Por consiguiente, podemos rechazar la hipótesis de medias iguales (es decir, no hay diferencia entre los tres tipos de tubos) al nivel de 0.05, pero no al nivel de 0.01.

9.10. Desarrolle el problema 9.9 usando las fórmulas abreviadas incluidas en (*24*), (*25*) y (*26*).

De la tabla 9-15,

$$n_1 = 3, \quad n_2 = 5, \quad n_3 = 4, \quad n = 12, \quad \tau_{1.} = 27, \quad \tau_{2.} = 25, \quad \tau_{3.} = 32, \quad \tau = 84$$

Por consiguiente tenemos

$$v = \sum_{j,k} x_{jk}^2 - \frac{\tau^2}{n} = 7^2 + 11^2 + \cdots + 6^2 + 8^2 - \frac{(84)^2}{12} = 72$$

$$v_b = \sum_j \frac{\tau_{j.}^2}{n_j} - \frac{\tau^2}{n} = \frac{(27)^2}{3} + \frac{(25)^2}{5} + \frac{(32)^2}{4} - \frac{(84)^2}{12} = 36$$

$$v_w = v - v_b = 36$$

Usando éstas, el análisis de varianza se lleva a cabo como en el problema 9.9.

CLASIFICACIÓN A DOS VÍAS O EXPERIMENTOS DE DOS FACTORES

9.11. La tabla 9-17 muestra la producción por hectárea de cuatro cultivos diferentes de plantas que crecen en lotes tratados con tres tipos distintos de fertilizante. Usando el método largo, pruebe al nivel de significancia de 0.01 si (*a*) hay diferencia significativa en la producción por hectárea debida a los fertilizantes, (*b*) hay diferencia significativa en la producción por hectárea debida al cultivo.

Tabla 9-17

	Cultivo I	Cultivo II	Cultivo III	Cultivo IV
Fertilizante A	4.5	6.4	7.2	6.7
Fertilizante B	8.8	7.8	9.6	7.0
Fertilizante C	5.9	6.8	5.7	5.2

Calcule los totales de las filas y las medias de las filas, así como los totales de las columnas y las medias de las columnas y la gran media, como se muestra en la tabla 9-18.

Tabla 9-18

	Cultivo I	Cultivo II	Cultivo III	Cultivo IV	Total de las filas	Media de las filas
Fertilizante A	4.5	6.4	7.2	6.7	24.8	6.2
Fertilizante B	8.8	7.8	9.6	7.0	33.2	8.3
Fertilizante C	5.9	6.8	5.7	5.2	23.6	5.9
Total de las columnas	19.2	21.0	22.5	18.9	Gran total = 81.6	
Media de las columnas	6.4	7.0	7.5	6.3	Gran media = 6.8	

ν_r = variación de las medias de las filas a partir de la gran media

$$= 4[(6.2 - 6.8)^2 + (8.3 - 6.8)^2 + (5.9 - 6.8)^2] = 13.68$$

ν_c = variación de las medias de las columnas a partir de la gran media

$$= 3[(6.4 - 6.8)^2 + (7.0 - 6.8)^2 + (7.5 - 6.8)^2 + (6.3 - 6.8)^2] = 2.82$$

ν = variación total

$$= (4.5 - 6.8)^2 + (6.4 - 6.8)^2 + (7.2 - 6.8)^2 + (6.7 - 6.8)^2$$
$$+ (8.8 - 6.8)^2 + (7.8 - 6.8)^2 + (9.6 - 6.8)^2 + (7.0 - 6.8)^2$$
$$+ (5.9 - 6.8)^2 + (6.8 - 6.8)^2 + (5.7 - 6.8)^2 + (5.2 - 6.8)^2$$
$$= 23.08$$

ν_e = variación aleatoria = $\nu - \nu_r - \nu_c = 6.58$.

Esto lleva al análisis de varianza de la tabla 9-19.

Al nivel de significancia de 0.05 con 2, 6 grados de libertad, $F_{0.95} = 5.14$. Entonces, ya que $6.24 > 5.14$, podemos rechazar la hipótesis de las medias de las filas son iguales y concluir, al nivel de significancia de 0.05, que hay una diferencia significativa debida a los fertilizantes.

Dado que el valor F correspondiente a las diferencias en las medias de las columnas es menor que 1, podemos concluir que no hay diferencia significativa en la producción debida a las plantas.

Tabla 9-19

Variación	Grados de libertad	Cuadrado medio	F
$v_r = 13.68$	2	$\hat{s}_r^2 = 6.84$	$F = \hat{s}_r^2/\hat{s}_e^2 = 6.24$ df: 2, 6
$v_c = 2.82$	3	$\hat{s}_c^2 = 0.94$	$F = \hat{s}_c^2/\hat{s}_e^2 = 0.86$ df: 3, 6
$v_e = 6.58$	6	$\hat{s}_e^2 = 1.097$	
$v = 23.08$	11		

9.12. Use las fórmulas de cálculo abreviadas para obtener los resultados del problema 9.11.

De la tabla 9-18 tenemos

$$\sum_{j,k} x_{jk}^2 = (4.5)^2 + (6.4)^2 + \cdots + (5.2)^2 = 577.96$$

$$\tau = 24.8 + 33.2 + 23.6 = 81.6$$

$$\sum \tau_{j.}^2 = (24.8)^2 + (33.2)^2 + (23.6)^2 = 2274.24$$

$$\sum \tau_{.k}^2 = (19.2)^2 + (21.0)^2 + (22.5)^2 + (18.9)^2 = 1673.10$$

Entonces

$$v = \sum_{j,k} x_{jk}^2 - \frac{\tau^2}{ab} = 577.96 - 554.88 = 23.08$$

$$v_r = \frac{1}{b}\sum \tau_{j.}^2 - \frac{\tau^2}{ab} = \frac{1}{4}(2274.24) - 554.88 = 13.68$$

$$v_c = \frac{1}{a}\sum \tau_{.k}^2 - \frac{\tau^2}{ab} = \frac{1}{3}(1673.10) - 554.88 = 2.82$$

$$v_e = v - v_r - v_c = 23.08 - 13.68 - 2.82 = 6.58$$

lo cual concuerda con el problema 9.11.

EXPERIMENTOS DE DOS FACTORES CON REPLICACIÓN

9.13. Un fabricante quiere determinar la efectividad de cuatro tipos de máquinas A, B, C, D en la producción de tornillos. Para lograrlo, se toma el número de tornillos defectuosos que produce cada máquina durante los días de cierta semana en los dos turnos de trabajo. Los resultados aparecen en la tabla 9-20. Lleve a cabo un análisis de varianza para probar con un nivel de significancia de 0.05 si hay (a) diferencia en las máquinas, (b) diferencia en los turnos.

Tabla 9-20

	Primer turno					Segundo turno				
	Lunes	Martes	Miércoles	Jueves	Viernes	Lunes	Martes	Miércoles	Jueves	Viernes
A	6	4	5	5	4	5	7	4	6	8
B	10	8	7	7	9	7	9	12	8	8
C	7	5	6	5	9	9	7	5	4	6
D	8	4	6	5	5	5	7	9	7	10

Los datos se pueden organizar de manera equivalente como en la tabla 9-21. En esta tabla se muestran los dos factores principales, a saber *máquina* y *turno*. Observe que para cada máquina se indican dos turnos. Los días de la semana se pueden considerar como réplicas o repeticiones del desempeño de cada máquina para los dos turnos.

Tabla 9-21

Factor I	Factor II	Réplicas					
Máquina	Turno	Lunes	Martes	Miércoles	Jueves	Viernes	Totales
A	1	6	4	5	5	4	24
	2	5	7	4	6	8	30
B	1	10	8	7	7	9	41
	2	7	9	12	8	8	44
C	1	7	5	6	5	9	32
	2	9	7	5	4	6	31
D	1	8	4	6	5	5	28
	2	5	7	9	7	10	38
	Totales	57	51	54	47	59	268

La variación total para todos los datos de la tabla 9-21 es

$$v = 6^2 + 4^2 + 5^2 + \cdots + 7^2 + 10^2 - \frac{(268)^2}{40} = 1946 - 1795.6 = 150.4$$

Con el fin de considerar los dos factores principales, *máquina* y *turno*, limitamos nuestra atención al total de valores de las réplicas correspondientes a cada combinación de estos factores. Éstos se acomodaron en la tabla 9-22, la cual es una tabla de dos factores con entradas sencillas.

Tabla 9-22

	Primer turno	Segundo turno	Totales
A	24	30	54
B	41	44	85
C	32	31	63
D	28	38	66
Totales	125	143	268

La variación total para la tabla 9-22, a la que podemos llamar *variación subtotal* v_s, está dada por

$$v_s = \frac{(24)^2}{5} + \frac{(41)^2}{5} + \frac{(32)^2}{5} + \frac{(28)^2}{5} + \frac{(30)^2}{5} + \frac{(44)^2}{5} + \frac{(31)^2}{5} + \frac{(38)^2}{5} - \frac{(268)^2}{40}$$
$$= 1861.2 - 1795.6 = 65.6$$

La variación entre filas está dada por

$$v_r = \frac{(54)^2}{10} + \frac{(85)^2}{10} + \frac{(63)^2}{10} + \frac{(66)^2}{10} - \frac{(268)^2}{40} = 1846.6 - 1795.6 = 51.0$$

La variación entre columnas está dada por

$$v_c = \frac{(125)^2}{20} + \frac{(143)^2}{20} - \frac{(268)^2}{40} = 1803.7 - 1795.6 = 8.1$$

Si ahora restamos de la variación subtotal v_s la suma de las variaciones entre filas y columnas $(\nu_r + \nu_s)$, obtenemos la variación debida a la interacción entre filas y columnas. Esto está dado por

$$v_i = v_s - v_r - v_c = 65.6 - 51.0 - 8.1 = 6.5$$

Finalmente, la variación residual, la cual podemos considerar como variación aleatoria o variación por error ν_e (siempre y cuando creamos que los distintos días de la semana no proveen diferencias importantes) se encuentra restando la suma de la fila, columna y variaciones por interacción (es decir, la variación subtotal) de la variación total ν. Esto da como resultado

$$v_e = v - (v_r + v_c + v_i) = v - v_s = 150.4 - 65.6 = 84.8$$

Estas variaciones se indican en el análisis de varianza, tabla 9-23. La tabla también da el número de grados de libertad correspondientes a cada tipo de variación. Por consiguiente, dado que hay 4 filas en la tabla 9-22, la variación debida a las filas tiene $4 - 1 = 3$ grados de libertad, mientras que la variación debida a las 2 columnas tiene $2 - 1 = 1$ grado de libertad. Para encontrar los grados de libertad debidos a las interacciones, observamos que en la tabla 9-22 hay 8 entradas. Por lo tanto, el total de grados de libertad es $8 - 1 = 7$. Dado que 3 de éstos se deben a las filas y 1 a las columnas, los restantes $7 - (3 + 1) = 3$, se deben a la interacción. Dado que hay 40 entradas en la tabla 9-21 original, el total de grados de libertad es $40 - 1 = 39$. Por consiguiente, los grados de libertad debidos a la variación aleatoria o residual son $39 - 7 = 32$.

Tabla 9-23

Variación	Grados de libertad	Cuadrado medio	F
Filas (máquinas), $v_r = 51.0$	3	$\hat{s}_r^2 = 17.0$	$\frac{17.0}{2.65} = 6.42$
Columnas (turnos), $v_c = 8.1$	1	$\hat{s}_c^2 = 8.1$	$\frac{8.1}{2.65} = 3.06$
Interacción $v_i = 6.5$	3	$\hat{s}_i^2 = 2.167$	$\frac{2.167}{2.65} = 0.817$
Subtotal, $v_s = 65.6$	7		
Aleatoria o residual, $v_e = 84.8$	32	$\hat{s}_e^2 = 2.65$	
Total, $v = 150.4$	39		

Para continuar debemos determinar primero si hay alguna interacción significativa entre los factores básicos (es decir, las filas y columnas de la tabla 9-22). De la tabla 9-23 vemos que para la interacción $F = 0.817$, lo cual muestra que la interacción no es significativa, o sea, no podemos rechazar la hipótesis $H_0^{(3)}$ de la página 345. Siguiendo las reglas de la página 345, vemos que el cálculo de F para las filas es 6.42. Dado que $F_{0.95} = 2.90$ para 3. 32 grados de libertad, podemos rechazar la hipótesis $H_0^{(1)}$ de que las filas tienen medias iguales. Esto equivale a decir que al nivel de 0.05 podemos concluir que las máquinas no son igualmente efectivas.

Para 1. 32 grados de libertad $F_{0.95} = 4.15$. Entonces, dado que el cálculo de F para las columnas es 3.06, *no* podemos rechazar la hipótesis $H_0^{(2)}$ de que las columnas tienen medias iguales. Esto equivale a decir que al nivel de 0.05 no hay diferencia significativa entre los turnos.

Si decidimos analizar los resultados combinando la interacción y las variaciones residuales como lo recomiendan algunos estadísticos, encontramos que para la variación combinada y los grados de libertad combinados $\nu_i + \nu_e = 6.5 + 84.8 = 91.3$ y $3 + 32 = 35$, respectivamente, lo cual nos lleva a la varianza combinada de $91.3/35 = 2.61$. Usar este valor en lugar de 2.65 para el denominador de F en la tabla 9-23 no afecta las conclusiones obtenidas anteriormente.

9.14. Desarrolle el problema 9.13 si se usa un nivel de significancia de 0.01.

A este nivel no hay todavía una interacción apreciable, así que podemos continuar.

Dado que $F_{0.99} = 4.47$ para 3.32 grados de libertad, y que el valor calculado de F para las filas es 6.42, podemos concluir que incluso al nivel de 0.01 las máquinas no son igualmente efectivas.

Dado que $F_{0.99} = 7.51$ para 1.32 grados de libertad, y que el valor calculado de F para las filas es 3.06, podemos concluir que incluso al nivel de 0.01 no hay diferencia significativa en los turnos.

CUADRADO LATINO

9.15. Un granjero quiere probar los efectos de cuatro fertilizantes diferentes A, B, C, D en la producción de trigo. Con el fin de eliminar fuentes de error debidas a la variabilidad en la fertilidad del suelo, usa los fertilizantes organizados en un cuadrado latino, como se muestra en la tabla 9-24, donde los números indican la producción en toneladas por unidad de área. Desarrolle un análisis de varianza para determinar si hay una diferencia significativa al nivel de (*a*) 0.05, (*b*) 0.01 entre los fertilizantes.

Tabla 9-24

A 18	*C* 21	*D* 25	*B* 11
D 22	*B* 12	*A* 15	*C* 19
B 15	*A* 20	*C* 23	*D* 24
C 22	*D* 21	*B* 10	*A* 17

Tabla 9-25

					Totales
	A 18	*C* 21	*D* 25	*B* 11	75
	D 22	*B* 12	*A* 15	*C* 19	68
	B 15	*A* 20	*C* 23	*D* 24	82
	C 22	*D* 21	*B* 10	*A* 17	70
Totales	77	74	73	71	295

Tabla 9-26

	A	*B*	*C*	*D*	
Total	70	48	85	92	295

Primero obtenemos los totales de las filas y las columnas, como se indica en la tabla 9-25. También obtenemo la producción total para cada uno de los fertilizantes, como se muestra en la tabla 9-26. La variación total y la variaciones para las filas, columnas y tratamientos se obtienen como de costumbre. Encontramos

$$\text{Variación total} = v = (18)^2 + (21)^2 + (25)^2 + \cdots + (10)^2 + (17)^2 - \frac{(295)^2}{16}$$
$$= 5769 - 5439.06 = 329.94$$

$$\text{Variación entre filas} = v_r = \frac{(75)^2}{4} + \frac{(68)^2}{4} + \frac{(82)^2}{4} + \frac{(70)^2}{4} - \frac{(295)^2}{16}$$
$$= 5468.25 - 5439.06 = 29.19$$

$$\text{Variación entre columnas} = v_c = \frac{(77)^2}{4} + \frac{(74)^2}{4} + \frac{(73)^2}{4} + \frac{(71)^2}{4} - \frac{(295)^2}{16}$$
$$= 5443.75 - 5439.06 = 4.69$$

$$\text{Variación entre tratamientos} = v_t = \frac{(70)^2}{4} + \frac{(48)^2}{4} + \frac{(85)^2}{4} + \frac{(92)^2}{4} - \frac{(295)^2}{16}$$
$$= 5723.25 - 5439.06 = 284.19$$

El análisis de varianza aparece en la tabla 9-27.

Tabla 9-27

Variación	Grados de libertad	Cuadrado medio	F
Filas, 29.19	3	9.73	4.92
Columnas, 4.69	3	1.563	0.79
Tratamientos, 284.19	3	94.73	47.9
Residuos, 11.87	6	1.978	
Total, 329.94	15		

(*a*) Dado que $F_{0.95,\,3,\,6} = 4.76$, podemos rechazar la hipótesis de que hay medias iguales de filas al nivel de 0.05. Se deduce que al nivel de 0.05 hay una diferencia en la fertilidad del suelo entre una fila y otra.
Dado que el valor F para las columnas es menor que 1, concluimos que no hay diferencia en la fertilidad del suelo en las columnas.
Dado que el valor F para los tratamientos es $47.9 > 4.76$, concluimos que hay una diferencia entre los fertilizantes.

(*b*) Dado que $F_{0.99,3,\,6} = 9.78$, aceptamos la hipótesis de que no hay diferencia en la fertilidad del suelo en las filas (o columnas) al nivel de significancia de 0.01. Sin embargo, aún debemos concluir que hay una diferencia entre los fertilizantes al nivel de 0.01.

CUADRADO GRECO LATINO

9.16. Es de interés determinar si hay alguna diferencia en el kilometraje por cada galón entre la gasolina de marca A, B, C, D. Diseñe un experimento usando cuatro conductores diferentes, cuatro automóviles diferentes y cuatro caminos.

Dado que está involucrado el mismo número de marcas de gasolina, conductores, automóviles y caminos, podemos usar el *cuadrado greco latino*. Supongamos que los carros se representan por las filas y los conductores por las columnas, como se indica en la tabla 9-28. Ahora, asignamos al azar las marcas de gasolina A, B, C, D a las filas y columnas, sujeto solamente a que cada letra aparezca sólo una vez en cada fila y en cada columna. Por consiguiente, cada conductor tendrá la oportunidad de manejar cada automóvil y usar cada tipo de gasolina (y no habrá un automóvil que se conduzca dos veces con la misma gasolina).

Ahora asignamos al azar los cuatro caminos que se usarán, denotándolos como α, β, γ, δ, y haciendo que cumplan las mismas restricciones impuestas en los cuadrados greco latinos. Por lo tanto, cada conductor tendrá la oportunidad de conducir en cada carretera. Uno de los posibles arreglos se da en la tabla 9-28.

Tabla 9-28

		Conductores			
		1	**2**	**3**	**4**
Automóviles	**1**	B_γ	A_β	D_δ	C_α
	2	A_δ	B_α	C_γ	D_β
	3	D_α	C_δ	B_β	A_γ
	4	C_β	D_γ	A_α	B_δ

9.17. Supongamos que la tabla 9.29 muestra los resultados del desarrollo del experimento del problema 9.16, el número de kilómetros por galón. Use el análisis de varianza para determinar si hay alguna diferencia significativa al nivel de 0.05.

Primero obtenemos los totales de las filas y las columnas, como se muestra en la tabla 9-30.

Tabla 9-29

Automóviles	Conductores 1	2	3	4
1	B_γ 19	A_β 16	D_δ 16	C_α 14
2	A_δ 15	B_α 18	C_γ 11	D_β 15
3	D_α 14	C_δ 11	B_β 21	A_γ 16
4	C_β 16	D_γ 16	A_α 15	B_δ 23

Tabla 9-30

					Totales
	B_γ 19	A_β 16	D_δ 16	C_α 14	65
	A_δ 15	B_α 18	C_γ 11	D_β 15	59
	D_α 14	C_δ 11	B_β 21	A_γ 16	62
	C_β 16	D_γ 16	A_α 15	B_δ 23	70
Totales	64	61	63	68	256

Luego, obtenemos los totales para cada letra latina y cada letra griega, así:

$$A \text{ total: } 15 + 16 + 15 + 16 = 62$$
$$B \text{ total: } 19 + 18 + 21 + 23 = 81$$
$$C \text{ total: } 16 + 11 + 11 + 14 = 52$$
$$D \text{ total: } 14 + 16 + 16 + 15 = 61$$
$$\alpha \text{ total: } 14 + 18 + 15 + 14 = 61$$
$$\beta \text{ total: } 16 + 16 + 21 + 15 = 68$$
$$\gamma \text{ total: } 19 + 16 + 11 + 16 = 62$$
$$\delta \text{ total: } 15 + 11 + 16 + 23 = 65$$

Ahora, calculamos las variaciones correspondientes a todas, usando el método abreviado.

$$\text{Filas: } \frac{(65)^2}{4} + \frac{(59)^2}{4} + \frac{(62)^2}{4} + \frac{(70)^2}{4} - \frac{(256)^2}{16} = 4112.50 - 4096 = 16.50$$

$$\text{Columnas: } \frac{(64)^2}{4} + \frac{(61)^2}{4} + \frac{(63)^2}{4} + \frac{(68)^2}{4} - \frac{(256)^2}{16} = 4102.50 - 4096 = 6.50$$

$$\text{Marcas de gasolina: } (A, B, C, D) \quad \frac{(62)^2}{4} + \frac{(81)^2}{4} + \frac{(52)^2}{4} + \frac{(61)^2}{4} - \frac{(256)^2}{16} = 4207.50 = 4096 = 111.50$$

$$\text{Caminos: } (\alpha, \beta, \gamma, \delta) \quad \frac{(61)^2}{4} + \frac{(68)^2}{4} + \frac{(62)^2}{4} + \frac{(65)^2}{4} - \frac{(256)^2}{16} = 4103.50 - 4096 = 7.50$$

La variación total es

$$(19)^2 + (16)^2 + (16)^2 + \cdots + (15)^2 + (23)^2 - \frac{(256)}{16} = 4244 - 4096 = 148.00$$

de manera que la variación debida al error es

$$148.00 - 16.50 - 6.50 - 111.50 - 7.50 = 6.00$$

Los resultados aparecen en el análisis de varianza, tabla 9-31. El número total de grados de libertad es $n^2 - 1$ para un *cuadrado de* $n \times n$. Las filas, las columnas, las letras latinas y las letras griegas tienen, cada una, $n - 1$ grados de libertad. Por consiguiente, los grados de libertad del error son $n^2 - 1 - 4(n - 1) = (n - 1)(n - 3)$. En nuestro caso, $n = 4$.

Tenemos $F_{0.95, 3, 3} = 9.28$ y $F_{0.99, 3, 3} = 29.5$. Por lo tanto, podemos rechazar la hipótesis de que las marcas de gasolina son iguales al nivel de 0.05, pero no al de 0.01.

Tabla 9-31

Variación	Grados de libertad	Cuadrado medio	F
Filas (automóviles), 16.50	3	5.500	$\frac{5.500}{2.000} = 2.75$
Columnas (conductores), 6.50	3	2.167	$\frac{2.167}{2.000} = 1.08$
Marcas de gasolina (A, B, C, D), 111.50	3	37.167	$\frac{37.167}{2.000} = 18.6$
Caminos ($\alpha, \beta, \gamma, \delta$), 7.50	3	2.500	$\frac{2.500}{2.000} = 1.25$
Error, 6.00	3	2.000	
Total, 148.00	15		

PROBLEMAS DIVERSOS

9.18. Demuestre que $\sum \alpha_j = 0$, [(*15*) página 337].

El tratamiento medias de la población viene dado por $\mu_j = \mu + \alpha_j$. Ya que,

$$\sum_{j=1}^{a} \mu_j = \sum_{j=1}^{a} \mu + \sum_{j=1}^{a} \alpha_j = a\mu + \sum_{j=1}^{a} \alpha_j = \sum_{j=1}^{a} \mu_j + \sum_{j=1}^{a} \alpha_j$$

donde usamos la definición $\mu = (\sum \mu_j)/a$. Se deduce que $\sum \alpha_j = 0$.

9.19. Derive (*a*) la ecuación (*17*), (*b*) la ecuación (*16*), de la página 338.

(*a*) Por definición tenemos

$$V_w = \sum_{j,k} (X_{jk} - \bar{X}_{j.})^2$$

$$= b \sum_{j=1}^{a} \left[\frac{1}{b} \sum_{k=1}^{b} (X_{jk} - \bar{X}_{j.})^2 \right]$$

$$= b \sum_{j=1}^{a} S_j^2$$

donde S_j^2 es la varianza muestral para el j-ésimo tratamiento, como se definió en (*15*), capítulo 5. Entonces, dado que el tamaño muestral es b,

$$E(V_w) = b \sum_{j=1}^{a} E(S_j^2)$$

$$= b \sum_{j=1}^{a} \left(\frac{b-1}{b} \sigma^2 \right)$$

$$= a(b-1)\sigma^2$$

usando (*16*) del capítulo 5.

(*b*) Por definición,

$$V_b = b\sum_{j=1}^{a}(\bar{X}_{j.} - \bar{X})^2$$

$$= b\sum_{j=1}^{a}\bar{X}_{j.}^2 - 2b\bar{X}\sum_{j=1}^{a}\bar{X}_{j.} + ab\bar{X}^2$$

$$= b\sum_{j=1}^{a}\bar{X}_{j.}^2 - ab\bar{X}^2$$

ya que

$$\bar{X} = \frac{\sum_j \bar{X}_{j.}^2}{a}$$

Entonces, omitiendo el índice de la sumatoria, tenemos

$$E(V_b) = b\sum E(\bar{X}_{j.}^2) - abE(\bar{X})^2 \tag{1}$$

Ahora para cualquier variable aleatoria U, $E(U^2) = \mathbf{Var}(U) + [E(U)]^2$. Por consiguiente,

$$E(\bar{X}_{j.}^2) = \mathbf{Var}(\bar{X}_{j.}) + [E(\bar{X}_{j.})]^2 \tag{2}$$

$$E(\bar{X}^2) = \mathbf{Var}(\bar{X}) + [E(\bar{X})]^2 \tag{3}$$

Pero, dado que las poblaciones de tratamiento tienen distribución normal con media μ_j y varianza común σ^2, según el teorema 5-4, página 168 tenemos:

$$\mathbf{Var}(\bar{X}_{j.}) = \frac{\sigma^2}{b} \tag{4}$$

$$\mathbf{Var}(\bar{X}) = \frac{\sigma^2}{ab} \tag{5}$$

$$E(\bar{X}_{j.}) = \mu_j = \mu + \alpha_j \tag{6}$$

$$E(\bar{X}) = \mu \tag{7}$$

Usando los resultados de (*2*) a (*7*), además del resultado del problema 9.18, en (*1*) tenemos

$$E(V_b) = b\sum\left[\frac{\sigma^2}{b} + (\mu + \alpha_j)^2\right] - ab\left[\frac{\sigma^2}{ab} + \mu^2\right]$$

$$= a\sigma^2 + b\sum(\mu + \alpha_j)^2 - \sigma^2 - ab\mu^2$$

$$= (a-1)\sigma^2 + ab\mu^2 + 2b\mu\sum\alpha_j + b\sum\alpha_j^2 - ab\mu^2$$

$$= (a-1)\sigma^2 + b\sum\alpha_j^2$$

9.20. Demuestre el teorema 9-1, página 338.

Como se muestra en el problema 9.19(*a*),

$$V_w = b\sum_{j=1}^{a}S_j^2 \qquad \text{o} \qquad \frac{V_w}{\sigma^2} = \sum_{j=1}^{a}\frac{bS_j^2}{\sigma^2}$$

donde S_j^2 es la varianza muestral para las muestras de tamaño b tomadas de la población de tratamiento j. Según el teorema 5-6, página 171, bS_j^2/σ^2 tiene distribución chi cuadrado con $b - 1$ grados de libertad. Entonces, ya que las varianzas S_j^2 son independientes, concluimos a partir del teorema 4-4, página 125, que V_w/σ^2 tiene distribución chi cuadrado con $a(b - 1)$ grados de libertad.

9.21. En el problema 9.13 supusimos que no había diferencia significativa en las réplicas, es decir, los diferentes días de la semana. ¿Podemos soportar esta conclusión al nivel de significancia de (*a*) 0.05, (*b*) 0.01?

Si existe alguna variación debida a las réplicas, está incluida en lo que se denomina "residuo" o "error aleatorio", $v_e = 8.48$, en la tabla 9-23. Para encontrar la variación debida a la réplica, usamos la columna de totales en la tabla 9-21, y obtenemos

$$v_{\text{rep}} = \frac{(57)^2}{8} + \frac{(51)^2}{8} + \frac{(54)^2}{8} + \frac{(47)^2}{8} + \frac{(59)^2}{8} - \frac{(268)^2}{40}$$
$$= 1807 - 1795.6 = 11.4$$

Dado que hay 5 réplicas, el número de grados de libertad asociado con esta variación es $5 - 1 = 4$. La variación residual luego de restar la variación debida a las réplicas es $v'_e = 84.8 - 11.4 = 73.4$. Las otras variaciones son las mismas de la tabla 9-23. La tabla de análisis de varianza final, teniendo en cuenta las réplicas, es la tabla 9-32.

Tabla 9-32

Variación	Grados de libertad	Cuadrado medio	F
Filas (máquinas), $v_r = 51.0$	3	17.0	$\frac{17.0}{2.621} = 6.49$
Columnas (turnos), $v_c = 8.1$	1	8.1	$\frac{8.1}{2.621} = 3.05$
Réplicas (días de la semana), $v_{\text{rep}} = 11.4$	4	2.85	$\frac{2.85}{2.621} = 1.09$
Interacción $v_i = 6.5$	3	2.167	$\frac{2.167}{2.621} = 0.827$
Aleatoriedad o residuo, $v'_e = 73.4$	28	2.621	
Total, $v = 150.4$	39		

De la tabla vemos que la F calculada para las réplicas es 1.09. Pero, dado que $F_{0.95} = 2.71$ para 4. 28 grados de libertad, podemos concluir que no hay variación significativa al nivel de 0.05 (y por consiguiente al nivel de 0.01) debido a las réplicas, es decir, los días de la semana no son significativos. Las conclusiones concernientes a *máquinas* y *turnos* son las mismas que se obtuvieron en el problema 9.13.

9.22. Describa cómo se puede usar el análisis de varianza en una clasificación tridimensional o en experimentos de tres factores (con entradas sencillas). Muestre la tabla de análisis de varianza que se usará en tal caso.

Suponemos que la clasificación se efectúa en A grupos, denotados como $A_1, \ldots, A_a$; B grupos, denotados como $B_1, \ldots, B_b$; y C grupos, denotados como $C_1, \ldots, C_c$. El valor de A_j, B_k y C_l se denota como x_{jkl}. El valor $\bar{x}_{jk.}$, por ejemplo, denota la media de los valores de la clase C cuando A_j y B_k se mantienen fijos. Significados similares se dan a $\bar{x}_{j.l}$ y $\bar{x}_{.kl}$. El valor $\bar{x}_{j..}$ es la media de los valores de las clases B y C cuando A_j está fija. Finalmente, $\bar{x}$ denota la gran media.

Habrá una variación total dada por

$$v = \sum_{j,k,l} (x_{jkl} - \bar{x})^2 \tag{1}$$

la cual se puede partir en siete variaciones, como se indica en la tabla 9-33. Estas variaciones son entre clases del mismo tipo y entre clases de tipos diferentes (*interacciones*). La interacción entre todas las clases se llama *variación residual* o *aleatoria*.

Las siete variaciones en las que se puede dividir (*1*) están dadas por

$$v = v_A + v_B + v_C + v_{AB} + v_{BC} + v_{CA} + v_{ABC}$$

donde

$$v_A = bc \sum_j (\bar{x}_{j..} - \bar{x})^2, \qquad v_B = ca \sum_k (\bar{x}_{.k.} - \bar{x})^2, \qquad v_C = ab \sum_l (\bar{x}_{..l} - \bar{x})^2$$

$$v_{AB} = c \sum_{j,k} (\bar{x}_{jk.} - \bar{x}_{j..} - \bar{x}_{.k.} + \bar{x})^2$$

$$v_{BC} = a \sum_{k,l} (\bar{x}_{.kl} - \bar{x}_{.k} - \bar{x}_{..l} + \bar{x})^2$$

$$v_{CA} = b \sum_{j,l} (\bar{x}_{j.l} - \bar{x}_{..l} - \bar{x}_{j..} + \bar{x})^2$$

$$v_{ABC} = \sum_{j,k,l} (x_{jkl} - \bar{x}_{jk.} - \bar{x}_{j.l} - \bar{x}_{.kl} + \bar{x}_{j..} + \bar{x}_{.k.} + \bar{x}_{..l} - \bar{x})^2$$

Tabla 9-33

Variación	Grados de libertad	Cuadro medio	F
v_A (Entre grupos A)	$a-1$	$\hat{s}_A^2 = \dfrac{v_A}{a-1}$	$\hat{s}_A^2/\hat{s}_{ABC}^2$ $a-1$, $(a-1)(b-1)(c-1)$ grados de libertad
v_B (Entre grupos B)	$b-1$	$\hat{s}_B^2 = \dfrac{v_B}{b-1}$	$\hat{s}_B^2/\hat{s}_{ABC}^2$ $b-1$, $(a-1)(b-1)(c-1)$ grados de libertad
v_C (Entre grupos C)	$c-1$	$\hat{s}_C^2 = \dfrac{v_C}{c-1}$	$\hat{s}_C^2/\hat{s}_{ABC}^2$ $c-1$, $(a-1)(b-1)(c-1)$ grados de libertad
v_{AB} (Entre grupos A y B)	$(a-1)(b-1)$	$\hat{s}_{AB}^2 = \dfrac{v_{AB}}{(a-1)(b-1)}$	$\hat{s}_{AB}^2/\hat{s}_{ABC}^2$ $(a-1)(b-1)$, $(a-1)(b-1)(c-1)$ grados de libertad
v_{BC} (Entre grupos B y C)	$(b-1)(c-1)$	$\hat{s}_{BC}^2 = \dfrac{v_{BC}}{(b-1)(c-1)}$	$\hat{s}_{BC}^2/\hat{s}_{ABC}^2$ $(b-1)(c-1)$, $(a-1)(b-1)(c-1)$ grados de libertad
v_{CA} (Entre grupos C y A)	$(c-1)(a-1)$	$\hat{s}_{CA}^2 = \dfrac{v_{CA}}{(c-1)(a-1)}$	$\hat{s}_{CA}^2/\hat{s}_{ABC}^2$ $(c-1)(a-1)$, $(a-1)(b-1)(c-1)$ grados de libertad
v_{ABC} (Entre grupos A, B y C)	$(a-1)(b-1)(c-1)$	$\hat{s}_{ABC}^2 = \dfrac{v_{ABC}}{(a-1)(b-1)(c-1)}$	
v (Total)	$abc-1$		

Problemas suplementarios

CLASIFICACIÓN A UNA VÍA O EXPERIMENTOS DE UN FACTOR

9.23. Se lleva a cabo un experimento para determinar la producción de 5 variedades de trigo *A*, *B*, *C*, *D*, *E*. A cada variedad se le asignan cuatro parcelas. La producción (en toneladas por hectárea) se muestra en la tabla 9-34. Suponiendo que las parcelas tienen fertilidad similar y las variedades se asignan a cada parcela al azar, determine si hay diferencia significativa en las producciones al nivel de significancia de (*a*) 0.05, (*b*) 0.01.

Tabla 9-34

A	20	12	15	19
B	17	14	12	15
C	23	16	18	14
D	15	17	20	12
E	21	14	17	18

Tabla 9-35

A	33	38	36	40	31	35
B	32	40	42	38	30	34
C	31	37	35	33	34	30
D	29	34	32	30	33	31

9.24. Una compañía quiere probar 4 tipos diferentes de llantas, *A*, *B*, *C*, *D*. Los tiempos de vida de éstas, determinados a partir de sus fibras, se dan (en miles de kilómetros) en la tabla 9-35, donde cada tipo se probó en 6 automóviles similares asignados al azar a las llantas. Pruebe a los niveles de (*a*) 0.05, (*b*) 0.01 si existe diferencia entre las llantas.

9.25. Un profesor quiere probar tres métodos diferentes de enseñar, I, II, III. Para hacerlo, se seleccionan al azar tres grupos de 5 estudiantes, y a cada grupo se lo enseña de manera diferente. Luego, se da a todos los estudiantes el mismo examen y se obtienen las notas que se muestran en la tabla 9-36. Determine al nivel de (*a*) 0.05, (*b*) 0.01 si hay diferencia significativa en los métodos de enseñanza.

Tabla 9-36

Método I	75	62	71	58	73
Método II	81	85	68	92	90
Método III	73	79	60	75	81

MODIFICACIONES PARA NÚMEROS DESIGUALES DE OBSERVACIONES

9.26. En la tabla 9-37 aparece el número de kilómetros por galón que se obtuvo cuando automóviles similares usaron 5 marcas diferentes de gasolina. Pruebe al nivel de significancia de (*a*) 0.05, (*b*) 0.01 si hay alguna diferencia significativa en las marcas.

Tabla 9-37

Marca *A*	12	15	14	11	15
Marca *B*	14	12	15		
Marca *C*	11	12	10	14	
Marca *D*	15	18	16	17	14
Marca *E*	10	12	14	12	

Tabla 9-38

Matemáticas	72	80	83	75	
Biología	81	74	77		
Inglés	88	82	90	87	80
Economía	74	71	77	70	

9.27. Durante un semestre un estudiante recibió las notas que aparecen en la tabla 9-38 en varias materias. Pruebe al nivel de (*a*) 0.05, (*b*) 0.01 si hay alguna diferencia significativa en las notas en estas materias.

CLASIFICACIÓN A DOS VÍAS O EXPERIMENTOS DE DOS FACTORES

9.28. Los artículos fabricados por una compañía son producidos por 3 operadores que usan 3 máquinas diferentes. El fabricante desea determinar si hay alguna diferencia (*a*) entre los operadores, (*b*) entre las máquinas. Un experimento se lleva a cabo para determinar el número de artículos por día que produce cada operador usando cada máquina. Los resultados se presentan en la tabla 9-39. Obtenga la información requerida usando un nivel de significancia de 0.05.

Tabla 9-39

	Operador 1	Operador 2	Operador 3
Máquina *A*	23	27	24
Máquina *B*	34	30	28
Máquina *C*	28	25	27

9.29. Desarrolle el problema 9.28 usando un nivel de significancia de 0.01.

9.30. En cinco bloques se sembraron semillas de cuatro tipos diferentes de maíz. Cada bloque se dividió en 4 parcelas, a las cuales se les asignó al azar los tipos de maíz. Pruebe al nivel de significancia de 0.05 si la producción en toneladas por hectárea, como se muestra en la tabla 9-40, varía significativamente con (*a*) las diferencias en el suelo (es decir, los 5 bloques), (*b*) las diferencias en el tipo de maíz.

Tabla 9-40

		Tipos de Maíz			
		I	*II*	*III*	*IV*
Bloques	*A*	12	15	10	14
	B	15	19	12	11
	C	14	18	15	12
	D	11	16	12	16
	E	16	17	11	14

9.31. Desarrolle el problema 9.30 usando un nivel de significancia de 0.01.

9.32. Suponga que la primera observación de cada tipo de llanta en el problema 9.24 se hace utilizando un tipo de automóvil en particular, la segunda observación usando un segundo tipo de automóvil, y así sucesivamente. Pruebe si hay alguna diferencia al nivel de 0.05 en (*a*) el tipo de llantas, (*b*) el tipo de automóvil.

9.33. Desarrolle el problema 9.32 usando un nivel de significancia de 0.01.

9.34. Suponga que en el problema 9.25 la primera entrada para cada método de enseñanza corresponde a un estudiante en un colegio en particular, el segundo método a otro estudiante en otro colegio, y así sucesivamente. Pruebe la hipótesis, al nivel del 0.05, de que hay una diferencia en (*a*) los métodos de enseñanza, (*b*) los colegios.

9.35. Se llevó a cabo un experimento para probar si el color del pelo y la estatura de las estudiantes adultas en Estados Unidos influye en su desempeño académico. Los resultados se presentan en la tabla 9-41, donde los números indican las estudiantes que se encuentran en el 10% superior de las graduandas. Analice el experimento al nivel de 0.05.

Tabla 9-41

	Pelirroja	Rubia	Castaña
Alta	75	78	80
Mediana	81	76	79
Baja	73	75	77

9.36. Desarrolle el problema 9.35 al nivel de 0.01.

EXPERIMENTOS DE DOS FACTORES CON REPLICACIÓN

9.37. Suponga que el experimento del problema 9.23 se llevó a cabo en el sur de Estados Unidos y que las columnas de la tabla 9-34 indican ahora 4 tipos diferentes de fertilizante, mientras que en un experimento similar, practicado en el oeste, se obtienen los resultados de la tabla 9-42. Pruebe al nivel de significancia de 0.05 si hay alguna diferencia entre (*a*) los fertilizantes, (*b*) la ubicación.

Tabla 9-42

A	16	18	20	23
B	15	17	16	19
C	21	19	18	21
D	18	22	21	23
E	17	18	24	20

9.38. Desarrolle el problema 9.37 usando un nivel de significancia de 0.01.

9.39. La tabla 9-43 da el número de artículos producidos por 4 operadores diferentes que trabajan en dos tipos de máquinas, I y II, en diferentes días de la semana. Determine al nivel de 0.05 si hay diferencias significativas en (*a*) los operadores, (*b*) las máquinas.

Tabla 9-43

	Máquina *I*					Máquina *II*				
	Lunes	Martes	Miércoles	Jueves	Viernes	Lunes	Martes	Miércoles	Jueves	Viernes
Operador *A*	15	18	17	20	12	14	16	18	17	15
Operador *B*	12	16	14	18	11	11	15	12	16	12
Operador *C*	14	17	18	16	13	12	14	16	14	11
Operador *D*	19	16	21	23	18	17	15	18	20	17

CUADRADO LATINO

9.40. Un experimento se lleva a cabo para probar el efecto de 4 fertilizantes diferentes *A*, *B*, *C*, *D*, sobre la producción de un campo de maíz, y las variaciones del suelo en dos sentidos perpendiculares. Se obtiene el cuadrado latino

de la tabla 9-44, en donde los números indican la producción de maíz por unidad de área. Pruebe al nivel de 0.01 la hipótesis de que no hay diferencia en (*a*) los fertilizantes, (*b*) las variaciones del suelo.

Tabla 9-44

C 8	*A* 10	*D* 12	*B* 11
A 14	*C* 12	*B* 11	*D* 15
D 10	*B* 14	*C* 16	*A* 10
B 7	*D* 16	*A* 14	*C* 12

9.41. Desarrolle el problema 9.40 usando un nivel de 0.05.

9.42. Con relación al problema 9.35, suponga que introducimos un factor adicional: anotar el lugar del país en donde nació el estudiante, *E*, *M* o *W*, tal como aparece en la tabla 9-45. Determine al nivel de 0.05 si hay una diferencia significativa en el desempeño académico de las estudiantes debido a las diferencias en (*a*) la estatura, (*b*) el color del pelo, (*c*) el lugar de nacimiento.

Tabla 9-45

E 75	*W* 78	*M* 80
M 81	*E* 76	*W* 79
W 73	*M* 75	*E* 77

CUADRADO GRECO LATINO

9.43. Con el fin de producir un tipo superior de alimento para pollos, se agregaron 4 cantidades diferentes de dos sustancias químicas a los ingredientes básicos. Las diferentes cantidades de la primera sustancia se indican como A, B, C, D, mientras que las de la segunda sustancia se indican como $\alpha, \beta, \gamma, \delta$. El alimento se da a los pollitos, los cuales están agrupados según 4 pesos iniciales W_1, W_2, W_3, W_4, y las 4 especies a las que pertenecen S_1, S_2, S_3, S_4. El incremento en el peso por unidad de tiempo se da en el cuadrado greco latino de la tabla 9-46. Desarrolle un análisis de varianza del experimento al nivel de significancia de 0.05, planteando toda conclusión que obtenga.

Tabla 9-46

	W_1	W_2	W_3	W_4
S_1	C_γ 8	B_β 6	A_α 5	D_δ 6
S_2	A_δ 4	D_α 3	C_β 7	B_γ 3
S_3	D_β 5	A_γ 6	B_δ 5	C_α 6
S_4	B_α 6	C_δ 10	D_γ 10	A_β 8

9.44. Cuatro compañías C_1, C_2, C_3, C_4 fabrican cuatro tipos diferentes de cables T_1, T_2, T_3, T_4. Cuatro operadores A, B, C, D miden la resistencia de los cables usando cuatro máquinas diferentes $\alpha, \beta, \gamma, \delta$. Las resistencias promedio obtenidas aparecen en el cuadrado greco latino de la tabla 9-47. Lleve a cabo un análisis de varianza al nivel de 0.05, planteando cualquier conclusión que obtenga.

Tabla 9-47

	C_1	C_2	C_3	C_4
T_1	A_β 164	B_γ 181	C_α 193	D_δ 160
T_2	C_δ 171	D_α 162	A_γ 183	B_β 145
T_3	D_γ 198	C_β 221	B_δ 207	A_α 188
T_4	B_α 157	A_δ 172	D_β 166	C_γ 136

PROBLEMAS DIVERSOS

9.45. La tabla 9-48 muestra los datos del óxido acumulado en el hierro tratado con los químicos A, B o C, respectivamente. Determine al nivel de significancia de (*a*) 0.05, (*b*) 0.01 si hay una diferencia significativa en los tratamientos.

Tabla 9-48

A	3	5	4	4
B	4	2	3	3
C	6	4	5	5

9.46. Un experimento mide el CI de estudiantes varones de estatura alta, baja y media. Los resultados aparecen en la tabla 9-49. Determine al nivel de significancia de (*a*) 0.05, (*b*) 0.01 si hay alguna diferencia significativa en los puntajes de CI relacionados con las diferentes estaturas.

Tabla 9-49

Alto	110	105	118	112	90	
Mediano	95	103	115	107		
Bajo	108	112	93	104	96	102

9.47. Se practica un examen para determinar si veteranos y no veteranos con diferentes CI se desempeñan mejor. Los puntajes obtenidos se muestran en la tabla 9-50. Determine al nivel de 0.05 si hay diferencia en el puntaje debida a diferencias en (*a*) el estatus de veterano, (*b*) CI.

Tabla 9-50

	CI alto	CI medio	CI bajo
Veterano	90	81	74
No veterano	85	78	70

9.48. Desarrolle el problema 9.47 usando un nivel de 0.01.

9.49. La tabla 9-51 muestra las calificaciones de una prueba practicada a una muestra de estudiantes universitarios de diferentes partes del país con diferentes CI. Analice la tabla al nivel de significancia de 0.05 y plantee sus conclusiones.

Tabla 9-51

	Alto	Medio	Bajo
Este	88	80	72
Oeste	84	78	75
Sur	86	82	70
Norte y centro	80	75	79

9.50. Desarrolle el problema 9.49 al nivel de 0.01.

9.51. Suponga que los resultados de la tabla 9-48 del problema 9.48 son válidos para la parte nororiental de Estados Unidos, mientras que los resultados correspondientes a la parte oeste se dan en la tabla 9.52. Determine al nivel de 0.05 si hay diferencias debidas a (*a*) los químicos, (*b*) la ubicación.

Tabla 9-52

A	5	4	6	3
B	3	4	2	3
C	5	7	4	6

9.52. Con relación a los problemas 9.23 y 9.37, suponga que se lleva a cabo un experimento adicional en la parte nororiental de Estados Unidos, cuyos resultados aparecen en la tabla 9-53. Pruebe al nivel de 0.05 si hay una diferencia en (*a*) los fertilizantes, (*b*) los tres lugares.

Tabla 9-53

A	17	14	18	12
B	20	10	20	15
C	18	15	16	17
D	12	11	14	11
E	15	12	19	14

9.53. Desarrolle el problema 9.52 con un nivel de significancia de 0.01.

9.54. Lleve a cabo un análisis de varianza del cuadrado latino de la tabla 9-54 al nivel de 0.05 y plantee conclusiones.

Tabla 9-54

	FACTOR 1		
FACTOR 2	*B* 16	*C* 21	*A* 15
	A 18	*B* 23	*C* 14
	C 15	*A* 18	*B* 12

9.55. Lleve a cabo un análisis de varianza del cuadrado greco latino de la tabla 9-55 al nivel de 0.05 y plantee conclusiones.

Tabla 9-55

		FACTOR 1			
FACTOR 2	A_γ 6	B_β 12	C_δ 4	D_α 18
	B_δ 3	A_α 8	D_γ 15	C_β 14
	D_β 15	C_γ 20	B_α 9	A_δ 5
	C_α 16	D_δ 6	A_β 17	B_γ 7

Respuestas a los problemas suplementarios

9.23. Hay una diferencia significativa a ambos niveles en la producción.

9.24. No hay una diferencia significativa en ambos niveles en las llantas.

9.25. Hay una diferencia significativa en el método de enseñanza al nivel de 0.05, pero no al nivel de 0.01.

9.26. Hay una diferencia significativa en las marcas al nivel de 0.05, pero no al nivel de 0.01.

9.27. Hay una diferencia significativa en sus notas a ambos niveles.

9.28. No hay una diferencia significativa en los operadores ni en las máquinas.

9.29. No hay una diferencia significativa en los operadores ni en las máquinas.

9.30. Hay una diferencia significativa al nivel de 0.05 en los tipos de maíz, pero no en los suelos.

9.31. No hay una diferencia significativa al nivel de 0.01 en el tipo de maíz ni en los suelos.

9.32. Hay una diferencia significativa al nivel de 0.05 tanto en las llantas como en los automóviles.

9.33. No hay una diferencia significativa al nivel de 0.01 en las llantas ni en los automóviles.

9.34. Hay una diferencia significativa al nivel de 0.05 en los métodos de enseñanza, pero no en los colegios.

9.35. No hay una diferencia significativa en el color del pelo ni en la estatura.

9.36. La misma respuesta que en el problema 9.35.

9.37. Hay una diferencia significativa al nivel de 0.05 en las ubicaciones, pero no en los fertilizantes.

9.38. No hay una diferencia significativa al nivel de 0.01 en las ubicaciones de los fertilizantes.

9.39. Hay una diferencia significativa en los operadores, pero no en las máquinas.

9.40. No hay una diferencia significativa en los fertilizantes ni en el suelo.

9.41. La misma respuesta que en el problema 9.40.

9.42. No hay una diferencia significativa en el rendimiento académico debido a las diferencias en la estatura, el color del pelo o el lugar de nacimiento.

9.43. Hay diferencias significativas en las especies y en las cantidades de la primera sustancia química, pero no otra diferencia significativa.

9.44. Hay diferencias significativas en los tipos de cables, pero no diferencias significativas en su resistencia debidas a los operadores, las máquinas o las compañías.

9.45. No hay una diferencia significativa en los tratamientos a ambos niveles.

9.46. No hay diferencia significativa en los puntajes a ambos niveles.

9.47. Hay diferencias significativas en los puntajes de las calificaciones debidas tanto al estatus de veterano como al CI, al nivel de 0.05.

9.48. Al nivel de 0.01, las diferencias en los puntajes de la prueba debidas al estatus de veterano no son significativas, pero las debidas al CI sí lo son.

9.49. No hay diferencias significativas en los puntajes de la prueba de los estudiantes de diferentes partes del país, pero sí en la prueba debida al CI.

9.50. La misma respuesta que en el problema 9.49.

9.51. Hay diferencia significativa al nivel de 0.05 debida a los químicos o a las ubicaciones.

9.52. Hay diferencias significativas debidas a las ubicaciones, pero no a los fertilizantes.

9.53. No hay diferencias significativas debidas a las ubicaciones o a los fertilizantes.

9.54. No hay diferencias significativas debidas al factor 1, 2 o a los tratamientos A, B, C.

9.55. No hay diferencias significativas debidas a factores o a tratamientos.

CAPÍTULO 10

Pruebas no paramétricas

INTRODUCCIÓN

La mayoría de las pruebas de hipótesis y significancia o reglas de decisión, consideradas en capítulos previos, requieren de varias suposiciones acerca de la distribución de la población de la que se toman las muestras. Por ejemplo, en el capítulo 5, con frecuencia se necesitaba que la distribucion de la población fuera normal o casi normal.

En la práctica surgen situaciones en las cuales tales suposiciones pueden no justificarse o en las que hay duda de que aplican, como en el caso en que una población puede estar altamente sesgada. Debido a esto, los estadísticos han ideado diferentes pruebas y métodos que son independientes de las distribuciones de la población y de sus parámetros asociados. Éstas se denominan *pruebas no paramétricas*.

Las pruebas no paramétricas se pueden utilizar como un método abreviado para remplazar pruebas más complejas. Son especialmente valiosas al tratar con datos no numéricos, tales como los que surgen cuando los consumidores ordenan cereales u otros productos en orden de preferencia.

LA PRUEBA DEL SIGNO

Consideremos la tabla 10-1, que muestra los números de tornillos defectuosos producidos por dos tipos diferentes de máquinas (I y II) en 12 días consecutivos y donde se supone que las máquinas tienen la misma producción por día. Queremos probar la hipótesis H_0 de que no hay diferencia entre las máquinas: que las diferencias observadas entre las máquinas en términos del número de tornillos defectuosos que producen se deben puramente al azar, lo que equivale a decir que las muestras provienen de la misma población.

Una prueba no paramétrica sencilla para el caso de tales muestras pareadas, es la *prueba del signo*. Esta prueba consiste en tomar la diferencia entre los números de tornillos defectuosos por día y escribir solamente

Tabla 10-1

Día	1	2	3	4	5	6	7	8	9	10	11	12
Máquina I	47	56	54	49	36	48	51	38	61	49	56	52
Máquina II	71	63	45	64	50	55	42	46	53	57	75	60

el signo de la diferencia. Por ejemplo, para el día 1 tenemos $47 - 71$, lo cual es negativo. De esta manera, obtenemos de la tabla 10-1 la secuencia de signos

$$- \quad - \quad + \quad - \quad - \quad - \quad + \quad - \quad + \quad - \quad - \quad - \qquad (1)$$

(es decir, 3 signos más y 9 signos menos). Ahora, si es igualmente probable obtener un 1 o un 2, podemos esperar obtener 6 de cada uno. La prueba de H_0 es entonces equivalente a la que prueba si una moneda es balanceada cuando en 12 lanzamientos se obtienen 3 caras s1d y 9 sellos s2d. Esto involucra la distribución binomial del capítulo 4. El problema 10.1 muestra que usando una prueba de dos colas de esta distribución al nivel de significancia de 0.05, no podemos rechazar H_0; es decir, a este nivel no hay diferencia entre las máquinas.

Comentario 1. Si algún día las máquinas producen el mismo número de tornillos defectuosos, la diferencia de *cero* aparecerá en la secuencia (*1*). En tal caso, podemos omitir los valores de esta muestra y usar 11 en lugar de 12 observaciones.

Comentario 2. También se puede usar una aproximación normal a la distribución binomial, usando una corrección para continuidad (*véase* el problema 10.2).

Aunque la prueba del signo es particularmente útil para las muestras pareadas, como en la tabla 10-1, también se puede usar para problemas que involucran muestras sencillas (*véanse* problemas 10.3 y 10.4).

LA PRUEBA *U* DE MANN-WHITNEY

Consideremos la tabla 10-2, que muestra las resistencias de los cables fabricados a partir de dos aleaciones diferentes, I y II. Queremos decidir si existe o no una diferencia entre las muestras o, lo que es equivalente, si provienen o no de la misma población. Aunque este problema se puede desarrollar usando la prueba t del capítulo 7, es muy útil la prueba no paramétrica llamada la *prueba U de Mann-Whitney*, o simplemente, la *prueba U*. Esta prueba consta de los siguientes pasos:

Tabla 10-2

Aleación I				Aleación II				
18.3	16.4	22.7	17.8	12.6	14.1	20.5	10.7	15.9
18.9	25.3	16.1	24.2	19.6	12.9	15.2	11.8	14.7

Paso 1. Organice todos los valores muestrales en serie desde el más pequeño hasta el más grande, y asígneles órdenes (en este caso de 1 a 18). Si dos o más valores muestrales son idénticos (o sea, hay *valores empatados*, o simplemente, *empates*), a dichos valores muestrales se les asigna un orden igual a la *media* de los órdenes que hubieran recibido de otra manera. Si la entrada 18.9 de la tabla 10-2, fuera 18.3, habrían dos valores idénticos 18.3 que ocuparían los órdenes 12 y 13 en la serie, de manera que el orden asignado a cada uno sería $(1/2)(12 + 13) = 12.5$.

Paso 2. Encuentre la suma de los órdenes para cada una de las muestras. Denote esas sumas como R_1 y R_2, donde N_1 y N_2 son los tamaños muestrales respectivos. Para conveniencia, escoja N_1 como el tamaño más pequeño si ellos son diferentes, de manera que $N_1 \leq N_2$. Una diferencia significativa entre las sumas de los órdenes R_1 y R_2 implica una diferencia significativa entre las muestras.

Paso 3. Para probar la diferencia entre las sumas de los órdenes, use el estadístico

$$U = N_1 N_2 + \frac{N_1(N_1 + 1)}{2} - R_1 \qquad (2)$$

correspondiente a la muestra 1. La distribución muestral de U es simétrica y tiene media y varianza dadas, respectivamente, por las fórmulas

$$\mu_U = \frac{N_1 N_2}{2} \quad ; \quad \sigma_U^2 = \frac{N_1 N_2 (N_1 + N_2 + 1)}{12} \qquad (3)$$

Si N_1 y N_2 son al menos iguales a 8, resulta entonces que U tiene distribución casi normal, de manera que

$$z = \frac{U - \mu_U}{\sigma_U} \tag{4}$$

tiene distribución normal con media 0 y varianza 1. Usando el apéndice C, podemos decidir si las muestras son significativamente diferentes. El problema 10.5 muestra que hay una diferencia significativa al nivel de 0.05 entre los cables.

Comentario 3. Un valor correspondiente a la muestra 2 está dado por el estadístico

$$U = N_1N_2 + \frac{N_2(N_2+1)}{2} - R_2 \tag{5}$$

y tiene la misma distribución muestral del estadístico (*2*), con media y varianza de la fórmula (*3*). El estadístico (*5*) se relaciona con el estadístico (*2*), si U_1 y U_2 son los valores correspondientes a los estadísticos (*2*) y (*5*), respectivamente, entonces tenemos el resultado

$$U_1 + U_2 = N_1N_2 \tag{6}$$

También tenemos

$$R_1 + R_2 = \frac{N(N+1)}{2} \tag{7}$$

donde $N = N_1 + N_2$. El resultado (*7*) provee la verificación para los cálculos.

Comentario 4. El estadístico U en la ecuación (*2*) es el número total de veces que los valores de la muestra 1 preceden los valores de la muestra 2 cuando todos los valores muestrales están arreglados en orden ascendente de magnitud. Esto provee un *método de cálculo* alterno para encontrar U.

LA PRUEBA *H* DE KRUSKAL-WALLIS

La prueba U es no paramétrica y se usa para decidir si dos muestras provienen o no de la misma población. La *prueba H de Kruskal-Wallis*, o la *prueba H,* es una generalización de esto para k muestras.

Esta prueba se puede describir así: supongamos que tenemos k muestras de tamaños $N_1, N_2, \ldots, N_k$, siendo el tamaño total de todas las muestras juntas $N = N_1 + N_2 + \cdots + N_k$. Supongamos, además, que los datos de todas las muestras juntas se ordenan y que las sumas de los órdenes para las k muestras son $R_1, R_2, \ldots, R_k$, respectivamente. Si definimos el estadístico

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{N_j} - 3(N+1) \tag{8}$$

entonces se puede demostrar que la distribución muestral de H tiene casi distribución chi cuadrado con $k - 1$ grados de libertad, siempre y cuando $N_1, N_2, \ldots, N_k$ sean al menos 5.

La prueba H ofrece un método no paramétrico en el *análisis de varianza* para clasificaciones simples o experimentos con un factor, y es posible hacer generalizaciones.

LA PRUEBA *H* CORREGIDA PARA EMPATES

En el caso en que haya muchos empates entre las observaciones en los datos de las muestras, el valor H dado por el estadístico (*8*) es menor de lo que debería ser. El valor corregido de H, denotado por H_c, se obtiene dividiendo el valor dado en el estadístico (*8*) por el factor de corrección

$$1 - \frac{\sum(T^3 - T)}{N^3 - N} \tag{9}$$

donde T es el número de empates correspondientes a cada observación y donde la suma se toma sobre todas las observaciones. Si no hubiera empates, entonces $T = 0$ y el factor (*9*) se reduce a 1, de manera que no es necesaria la corrección. En la práctica, la corrección es por lo general despreciable (es decir, no es suficiente para garantizar un cambio en la decisión).

LA PRUEBA DE CORRIDAS PARA ALEATORIEDAD

A pesar de que la palabra *aleatorio* se ha utilizado muchas veces en este libro (tal como en muestreo aleatorio y lanzar una moneda al azar), ningún capítulo dio una prueba para aleatoriedad. La *teoría de corridas* da una prueba no paramétrica para la aleatoriedad.

Para entender lo que es una corrida, consideremos la secuencia hecha con dos símbolos, a y b, tal como

$$a\ a \mid b\ b\ b \mid a \mid b\ b \mid a\ a\ a\ a\ a \mid b\ b\ b \mid a\ a\ a\ a \mid \qquad (10)$$

En el lanzamiento de una moneda, por ejemplo, a puede representar las caras y b, los sellos. O en el muestreo de tornillos producidos por una máquina, a puede representar los defectuosos y b, los no defectuosos.

Una *corrida* se define como un conjunto de símbolos idénticos (o relacionados) contenido entre dos símbolos diferentes o no símbolos (como el inicio o final de una secuencia). Yendo de izquierda a derecha en la secuencia (*10*), la primera corrida, indicada por una barra vertical, consta de dos letras a; de igual manera, la segunda corrida consta de tres letras b, la tercera de una letra a, etc. Hay siete corridas en (*10*).

Parece claro que existe alguna relación entre la aleatoriedad y el número de corridas. Así, para la secuencia

$$a \mid b \mid a \mid b \mid a \mid b \mid a \mid b \mid a \mid b \mid a \mid b \mid \qquad (11)$$

hay un *patrón cíclico*, en el cual vamos de a hasta b, de nuevo a a, etc., lo cual difícilmente puede creerse que es aleatorio. En tal caso tenemos *demasiadas* corridas (de hecho, tenemos el número máximo posible para el número dado de a y b).

Por otro lado, en la secuencia

$$a\ a\ a\ a\ a\ a \mid b\ b\ b\ b \mid a\ a\ a\ a\ a \mid b\ b\ b \mid \qquad (12)$$

parece que hay un *patrón disperso*, en el cual las letras a y b se agrupan (o amontonan). En tal caso decimos que hay *muy pocas* corridas, y no consideraremos la secuencia como aleatoria.

Por lo tanto, una secuencia se puede considerar no aleatoria si hay muy pocas o demasiadas corridas, y se considera aleatoria de otra manera. Para cuantificar este concepto, supongamos que formamos todas las secuencias posibles que constan de N_1 letras a, N_2 letras b, para un total de N símbolos en todas ($N_1 + N_2 = N$). La colección de estas secuencias nos provee una distribución muestral. Cada secuencia tiene un número asociado de corridas, denotadas por V. De esta manera, llegamos a la distribución muestral del estadístico V. Se puede demostrar que esta distribución muestral tiene media y varianza dadas, respectivamente, mediante las fórmulas

$$\mu_V = \frac{2N_1N_2}{N_1 + N_2} + 1 \qquad \sigma_V^2 = \frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1 + N_2)^2(N_1 + N_2 - 1)} \qquad (13)$$

Usando las fórmulas (*13*) podemos probar la hipótesis de aleatoriedad a niveles apropiados de significancia. Resulta que si tanto N_1 como N_2 son, al menos, iguales a 8, entonces la distribución muestral de V es casi una distribución normal. Así que

$$z = \frac{V - \mu_V}{\sigma_V} \qquad (14)$$

tiene distribución normal con media 0 y varianza 1, y por lo tanto se puede usar el apéndice C.

APLICACIONES ADICIONALES DE LA PRUEBA DE CORRIDAS

Las siguientes son otras aplicaciones de la prueba de corridas para problemas estadísticos.

1. **PRUEBA DE VALORES POR ENCIMA O POR DEBAJO DE LA MEDIANA PARA ALEATORIEDAD DE DATOS NUMÉRICOS.** Para determinar si los datos numéricos (tales como los obtenidos en una muestra) son aleatorios, primero coloque los datos en el *mismo orden* en que fueron obtenidos. Luego, encuentre la mediana de ellos y remplace cada entrada con la letra a o b según si el valor está por encima o por debajo de la mediana. Si un valor es igual a la mediana, sáquelo de la muestra. La muestra es no aleatoria o si la secuencia de a y b es o no aleatoria (*véase* problema 10.20.)

2. **DIFERENCIAS EN POBLACIONES DE LAS QUE SE TOMAN MUESTRAS.** Supongamos que dos muestras de tamaño m y n se denotan como $a_1, a_2, \ldots, a_m$ y $b_1, b_2, \ldots, b_n$, respectivamente. Para decidir si las muestras provienen o no de la misma población, primero coloque todos los $m + n$ valores muestrales en una secuencia de valores en orden ascendente. Si algunos valores son los mismos, se deben ordenar en un proceso aleatorio (tal como usando números aleatorios). Si la secuencia resultante es aleatoria, podemos concluir que las muestras realmente no son diferentes, y por lo tanto, provienen de la misma población. Si la secuencia no es aleatoria, no podemos llegar a tal conclusión. Esta prueba provee una alternativa a la prueba U de Mann-Whitney (*véase* el problema 10.21).

CORRELACIÓN DE RANGO DE SPEARMAN

Los métodos no paramétricos también pueden usarse para medir la correlación de dos variables X y Y. En lugar de usar valores exactos de las variables, o cuando tal precisión no está disponible, los datos se deben colocar de 1 a N en orden de tamaño, importancia, etc. Si X y Y se colocan de tal manera, el *coeficiente de correlación de rango* o la *fórmula de Spearman para la correlación de rango* está dada por

$$r_S = 1 - \frac{6\sum D^2}{N(N^2 - 1)} \tag{15}$$

donde D denota las diferencias entre los órdenes de los valores X y Y correspondientes, y N es el número de pares de valores sX y Yd en los datos.

Problemas resueltos

LA PRUEBA DEL SIGNO

10.1. Con relación a la tabla 10-1, pruebe la hipótesis H_0 de que no hay diferencia entre las máquinas I y II, contra la hipótesis H_1 de que hay una diferencia al nivel de significancia de 0.05.

La figura 10-1 es una gráfica de la distribución binomial (y de la aproximación normal a ella) que da las probabilidades de X caras en 12 lanzamientos de una moneda balanceada, donde $X = 0, 1, 2, \ldots, 12$. Del capítulo 4, la probabilidad de X caras es

$$\Pr\{X\} = \binom{12}{X}\left(\frac{1}{2}\right)^X\left(\frac{1}{2}\right)^{12-X} = \binom{12}{X}\left(\frac{1}{2}\right)^{12}$$

de donde $\Pr\{0\} = 0.00024$, $\Pr\{1\} = 0.00293$, $\Pr\{2\} = 0.01611$ y $\Pr\{3\} = 0.05371$.

Dado que H_1 es la hipótesis de que hay una *diferencia* entre las máquinas, en contra de la hipótesis de que la máquina I es *mejor* que la máquina II, usamos una prueba de dos colas. Para el nivel de significancia de 0.05,

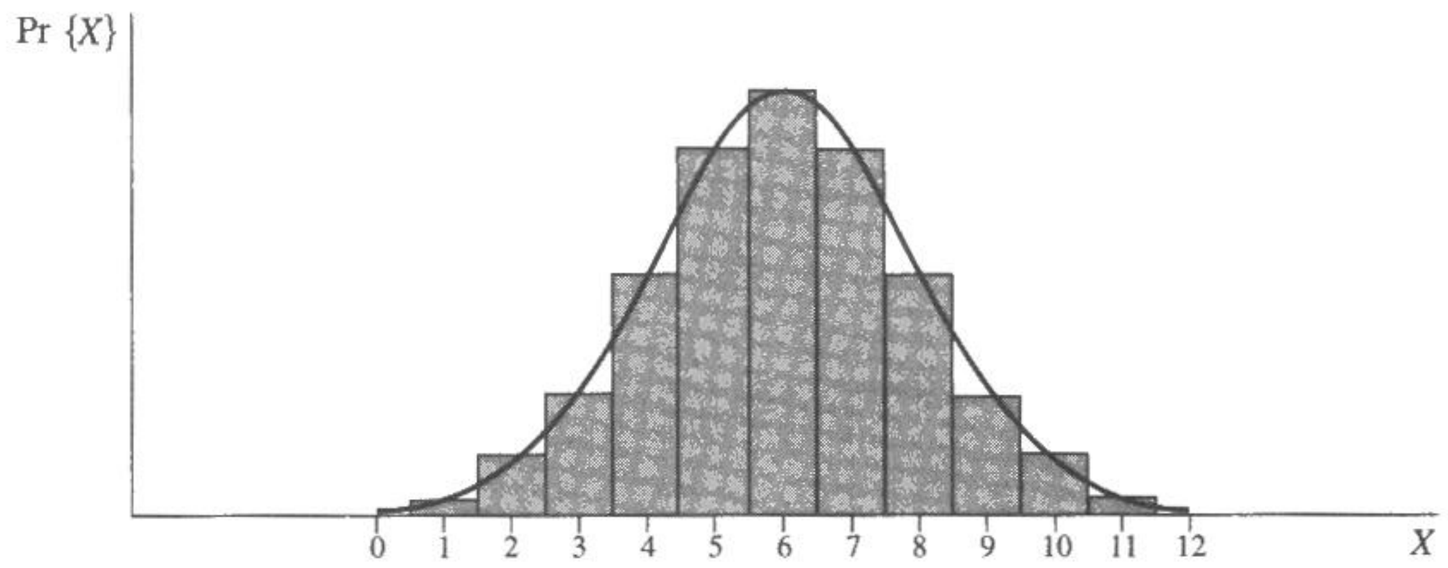


Fig. 10-1

cada cola tiene la probabilidad asociada $(\frac{1}{2})(0.05) = 0.025$. Ahora agregamos las probabilidades al lado izquierdo hasta que la suma supere 0.025. Así

$$\Pr\{0, 1, \text{ o } 2 \text{ caras}\} = 0.00024 + 0.00293 + 0.01611 = 0.01928$$

$$\Pr\{0, 1, 2, \text{ o } 3 \text{ caras}\} = 0.00024 + 0.00293 + 0.01611 + 0.05371 = 0.07299$$

Dado que 0.025 es mayor que 0.01928, pero menor que 0.07299, podemos rechazar la hipótesis H_0 si el número de caras es 2 o menor (o, por simetría, si el número de caras es 10 o mayor). Sin embargo, el número de caras [los signos + de la secuencia (*1*) de este capítulo] es 3. Por lo tanto, no podemos rechazar H_0 al nivel de 0.05 y debemos concluir que no hay diferencia entre las máquinas a este nivel.

10.2. Desarrolle el problema 10.1 usando una aproximación normal a la distribución binomial.

Para una aproximación normal a la distribución binomial, usamos el hecho de que el valor z correspondiente al número de caras es

$$z = \frac{X - \mu}{\sigma} = \frac{X - Np}{\sqrt{Npq}}.$$

Debido a que la variable X de la distribución binomial es discreta, mientras que para una distribución normal es continua, hacemos una *corrección por continuidad* (por ejemplo, 3 caras es un valor real entre 2.5 y 3.5 caras). Esto significa disminuir X por 0.5 si $X > Np$ e incrementar X por 0.5 si $X < Np$. Ahora $N = 12$, $\mu = Np = (12)(0.5) = 6$ y $\sigma = \sqrt{Npq} = \sqrt{(12)(0.5)(0.5)} = 1.73$ así que

$$z = \frac{(3 + 0.5) - 6}{1.73} = -1.45$$

Ya que esto es mayor que -1.96 (el valor de z para el cual el área de la izquierda es 0.025), llegamos a la misma conclusión que en el problema 10.1.

Observe que $Pr\{z \leq -1.45\} = 0.0735$, lo cual concuerda muy bien con el $Pr\{X \leq 3 \text{ caras}\} = 0.07299$ del problema 10.1.

10.3. El fabricante PQR sostiene que el tiempo de vida de un tipo de batería que se fabrican es de más de 250 horas. Un defensor de los consumidores, que desea determinar si la afirmación es justificada, mide los tiempos de vida media de 24 de las baterías de la compañía. Los resultados aparecen en la tabla 10-3. Suponiendo que las muestras son aleatorias, determine si la afirmación del fabricante es justificada al nivel de 0.05.

Tabla 10-3

271	230	198	275	282	225	284	219
253	216	262	288	236	291	253	224
264	295	211	252	294	243	272	268

Sea H_0 la hipótesis de que las baterías del fabricante tienen un tiempo de vida igual a 250 horas, y sea H_1 la hipótesis de que tienen un tiempo de vida mayor que 250 horas. Para probar H_0 en contra de H_1, usamos la prueba del signo. Para hacerlo, restamos 250 a cada dato en la tabla 10-3 y registramos los signos de las diferencias, como aparece en la tabla 10-4. Vemos que hay 15 signos más y 9 signos menos.

Tabla 10-4

+	−	−	+	+	−	+	−
+	−	+	+	−	+	+	−
+	+	−	+	+	−	+	+

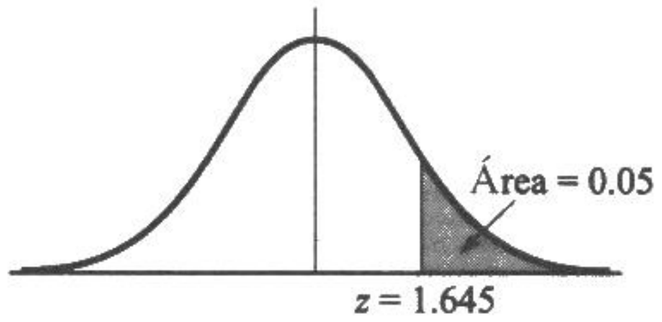


Fig. 10-2

Usando una prueba de una cola al nivel de significancia de 0.05, podemos rechazar H_0 si el valor z fuera mayor que 1.645 (figura 10-2). Dado que el valor z, usando la corrección por continuidad, es

$$z = \frac{(15 - 0.5) - (24)(0.5)}{\sqrt{(24)(0.5)(0.5)}} = 1.02$$

la afirmación del fabricante no se puede justificar al nivel de 0.05.

10.4. En la tabla 10-5 aparece la muestra de 40 notas de una prueba practicada en todo el país. Al nivel de 0.05 pruebe la hipótesis de que la nota promedio para todos los participantes es (*a*) 66, (*b*) 75.

Tabla 10-5

71	67	55	64	82	66	74	58	79	61
78	46	84	93	72	54	78	86	48	52
67	95	70	43	70	73	57	64	60	83
73	40	78	70	64	86	76	62	95	66

Tabla 10-6

+	+	−	−	+	0	+	−	+	−
+	−	+	+	+	−	+	+	−	−
+	+	+	−	+	+	−	−	−	+
+	−	+	+	−	+	+	−	+	0

(*a*) Restando 66 a todos los datos de la tabla 10-5 y reteniendo sólo los signos asociados, obtenemos la tabla 10-6, en la que vemos que hay 23 signos más, 15 signos menos y 2 ceros. Descartando los 2 ceros, nuestra muestra consta de 38 signos: 23 signos más y 15 signos menos. Usando una prueba de dos colas de la distribución normal con probabilidades $(\frac{1}{2})(0.05) = 0.025$ en cada cola (figura 10-3), adoptamos la siguiente regla de decisión:

Aceptar la hipótesis si $-1.96 \leq z \leq 1.96$.
Rechazar la hipótesis de otra manera.

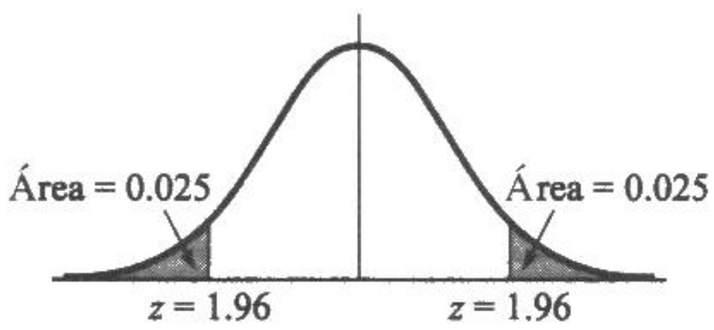


Fig. 10-3

Dado que $$z = \frac{X - Np}{\sqrt{Npq}} = \frac{(23 - 0.5) - (38)(0.5)}{\sqrt{(38)(0.5)(0.5)}} = 1.14$$

aceptamos la hipótesis de que la media es 66 al nivel de 0.05.

Observe que también hubiéramos podido usar 15, el número de signos menos. En este caso

$$z = \frac{(15 + 0.5) - (38)(0.5)}{\sqrt{(38)(0.5)(0.5)}} = -1.14$$

con la misma conclusión.

(*b*) Restamos 75 de todos los datos de la tabla 10-7, en la que hay 13 signos más y 27 signos menos. Dado que

$$z = \frac{(13 + 0.5) - (40)(0.5)}{\sqrt{(40)(0.5)(0.5)}} = -2.06$$

rechazamos la hipótesis de que la media es 75, al nivel de 0.05.

Tabla 10-7

−	−	−	−	+	−	−	−	+	−
+	−	+	+	−	−	+	+	−	−
−	+	−	−	−	−	−	−	−	+
−	−	+	−	−	+	+	−	+	−

Usando este método podemos llegar al intervalo de confianza del 95% para la mediana de las notas del examen (*véase* el problema 10.30).

LA PRUEBA *U* DE MANN-WHITNEY

10.5. Con relación a la tabla 10-2, determine si hay una diferencia al nivel de significancia de 0.05 entre los cables hechos con la aleación I y la aleación II.

Organizamos el trabajo según los pasos 1, 2 y 3 (descritos anteriormente en este capítulo):

Paso 1: La combinación de los valores de las 18 muestras en orden ascendente nos da la primera línea de la tabla 10-8. Estos valores están numerados de 1 a 18 en la segunda línea, que nos da los rangos.

Paso 2: Para encontrar la suma de los rangos para cada muestra, reescribimos la tabla 10-2 usando los rangos asociados de la tabla 10-8. Esto nos da la tabla 10-9. La suma de los rangos es 106 para la aleación I y 65 para la aleación II.

Tabla 10-8

10.7	11.8	12.6	12.9	14.1	14.7	15.2	15.9	16.1	16.4	17.8	18.3	18.9	19.6	20.5	22.7	24.2	25.3
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

Paso 3: Ya que la muestra de la aleación I tiene el tamaño más pequeño, $N_1 = 8$ y $N_2 = 10$, y las sumas correspondientes de los rangos son $R_1 = 106$ y $R_2 = 65$, entonces

$$U = N_1N_2 + \frac{N_1(N_1 + 1)}{2} - R_1 = (8)(10) + \frac{(8)(9)}{2} - 106 = 10$$

$$\mu_U = \frac{N_1N_2}{2} = \frac{(8)(10)}{2} = 40 \qquad \sigma_U^2 = \frac{N_1N_2(N_1 + N_2 + 1)}{12} = \frac{(8)(10)(19)}{12} = 126.67$$

Así $\sigma_U = 11.25$ y

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{10 - 40}{11.25} = -2.67$$

Tabla 10-9

Aleación I		Aleación II	
Resistencia del cable	Rango	Resistencia del cable	Rango
18.3	12	12.6	3
16.4	10	14.1	5
22.7	16	20.5	15
17.8	11	10.7	1
18.9	13	15.9	8
25.3	18	19.6	14
16.1	9	12.9	4
24.2	17	15.2	7
Suma	106	11.8	2
		14.7	6
		Suma	65

Dado que la hipótesis H_0 que estamos probando es si *no* existe diferencia entre las aleaciones, se requiere una prueba de dos colas. Para el nivel de significancia de 0.05, tenemos la siguiente regla de decisión:

Aceptar la H_0 si $-1.96 \leq z \leq 1.96$.
Rechazar la H_0 de otra manera.

Debido a que $z = -2.67$, rechazamos H_0 y concluimos que sí hay diferencia entre las aleaciones al nivel de 0.05.

10.6. Verifique los resultados (*6*) y (*7*) de este capítulo para los datos del problema 10.5.

(*a*) Ya que las muestras 1 y 2 dan valores para U dados por

$$U_1 = N_1N_2 + \frac{N_1(N_1+1)}{2} - R_1 = (8)(10) + \frac{(8)(9)}{2} - 106 = 10$$

$$U_2 = N_1N_2 + \frac{N_2(N_2+1)}{2} - R_2 = (8)(10) + \frac{(10)(11)}{2} - 65 = 70$$

tenemos $U_1 + U_2 = 10 + 70 = 80$, y $N_1N_2 = 106 + 65 = 171$ y

$$\frac{N(N+1)}{2} = \frac{(N_1+N_2)(N_1+N_2+1)}{2} = \frac{(18)(19)}{2} = 171$$

10.7. Desarrolle el problema 10.5 usando el estadístico U para la muestra de la aleación II.

Para la muestra de la aleación II,

$$U = N_1N_2 + \frac{N_2(N_2+1)}{2} - R_2 = (8)(10) + \frac{(10)(11)}{2} - 65 = 70$$

así que

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{70 - 40}{11.25} = 2.67$$

Este valor de z es el *negativo* de z en el problema 10.5, y se usa la cola del lado derecho de la distribución normal en lugar de la del lado izquierdo. Dado que este valor de z está por fuera de $-1.96 \leq z \leq 1.96$, la conclusión es la misma que la del problema 10.5.

10.8. Un profesor dicta dos clases en psicología: una en la mañana, con 9 estudiantes, y una en la tarde con 12 estudiantes. Todos los estudiantes presentaron el examen final a la misma hora, y las notas se muestran en la tabla 10-10. ¿Podemos concluir al nivel de significancia de 0.05 que el desempeño de la clase de la mañana es peor que el de la clase de la tarde?

Tabla 10-10

Clase de la mañana	73	87	79	75	82	66	95	75	70			
Clase de la tarde	86	81	84	88	90	85	84	92	83	91	53	84

Paso 1: La tabla 10-11 muestra el arreglo de las notas y los rangos. Observe que el rango de las dos notas de 75 es $(\frac{1}{2})(5 + 6) = 5.5$, mientras que el rango para las tres notas de 84 es $(\frac{1}{3})(11 + 12 + 13) = 12$.

Paso 2: Reescribimos la tabla 10-10 en términos de los rangos dados en la tabla 10-12.

Verificación: $R_1 = 73$, $R_2 = 158$ y $N = N_1 + N_2 = 9 + 12 = 21$. Así, $R_1 + R_2 = 73 + 158 = 231$ y

$$\frac{N(N+1)}{2} = \frac{(21)(22)}{2} = 231 = R_1 + R_2$$

Tabla 10-11

53	66	70	73	75	75	79	81	82	83	84	84	84	85	86	87	88	90	91	92	95
1	2	3	4	5.5		7	8	9	10		12		14	15	16	17	18	19	20	21

Tabla 10-12

													Suma de los rangos
Clases de la mañana	4	16	7	5.5	9	2	21	5.5	3				73
Clases de la tarde	15	8	12	17	18	14	12	20	10	19	1	12	158

Paso 3:

$$U = N_1N_2 + \frac{N_1(N_1+1)}{2} - R_1 = (9)(12) + \frac{(9)(10)}{2} - 73 = 80$$

$$\mu_U = \frac{N_1N_2}{2} = \frac{(9)(12)}{2} = 54 \quad ; \quad \sigma_U^2 = \frac{N_1N_2(N_1+N_2+1)}{12} = \frac{(9)(12)(22)}{12} = 198$$

Por consiguiente,
$$z = \frac{U - \mu_U}{\sigma_U} = \frac{80 - 54}{14.07} = -1.85$$

Dado que deseamos probar la hipótesis H_1 de que la clase de la mañana se desempeña peor que la de la tarde, en contra de la hipótesis H_0 de que no hay diferencia al nivel de 0.05, se necesita una prueba de una cola. Con relación a la figura 10.2, que aplica aquí, tenemos la regla de decisión:

Aceptar H_0 si $z \leq 1.645$.
Rechazar H_0 si $z > 1.645$.

Dado que el valor real de z es $z = 1.85 > 1.645$, rechazamos H_0 y concluimos al nivel de 0.05 que la clase de la mañana se desempeña peor que la de la tarde. Sin embargo, no se puede llegar a esta conclusión al nivel de 0.01 (*véase* el problema 10.33).

10.9. Encuentre U para los datos de la tabla 10-13 usando (a) la fórmula (2) de este capítulo, (b) el método de conteo (como se describió en el comentario 4 de este capítulo).

(a) Acomodando los datos de ambas muestras en orden ascendente de magnitud y asignando rangos de 1 a 5, obtenemos la tabla 10-14. Remplazando los datos de la tabla 10-13 con los rangos correspondientes, obtenemos la tabla 10-15, a partir de la cual las sumas de los rangos son $R_1 = 5$ y $R_2 = 10$. Ya que $N_1 = 2$ y $N_2 = 3$, el valor de U para la muestra 1 es

$$U = N_1N_2 + \frac{N_1(N_1+1)}{2} - R_1 = (2)(3) + \frac{(2)(3)}{2} - 5 = 4$$

El valor U para la muestra 2 se puede encontrar de manera similar siendo $U = 2$.

Tabla 10-13

Muestra 1	22	10	
Muestra 2	17	25	14

Tabla 10-14

Datos	10	14	17	22	25
Rangos	1	2	3	4	5

Tabla 10-15

				Suma de los rangos
Muestra 1	4	1		5
Muestra 2	3	5	2	10

(b) Remplacemos los valores muestrales de la tabla 10-14 con I o II, dependiendo de si el valor pertenece a la muestra 1 o a la 2. Entonces la primera línea de la tabla 10-14 se convierte en

Datos	I	II	II	I	II

A partir de esto vemos que

Número de valores de la muestra 1 que preceden el primer valor de la muestra 2 = 1
Número de valores de la muestra 1 que preceden el segundo valor de la muestra 2 = 1
Número de valores de la muestra 1 que preceden al tercer valor de la muestra 2 = 2
Total = 4

Así, el valor correspondiente de U para la primera muestra es 4.

Similarmente tenemos:

Número de valores de la muestra 2 que preceden el primer valor de la muestra 1 = 0
Número de valores de la muestra 2 que preceden el segundo valor de la muestra 1 = 2
Total = 2

Así, el valor de U correspondiente al segundo valor de la muestra 2 es 2.

Observe que dado que $N_1 = 2$ y $N_2 = 3$, estos valores satisfacen $U_1 + U_2 = N_1N_2$; que es, $4 + 2 = (2)(3) = 6$.

10.10. Una población consta de los valores 7, 12 y 15. Se toman dos muestras sin remplazo de esta población. La muestra 1 tiene un valor y la muestra 2 tiene dos valores [entre ellos, las dos muestras consumen toda la población].

(a) Encuentre la distribución muestral de U.

(*b*) Encuentre la media y la varianza de la distribución de la parte (*a*).

(*c*) Verifique los resultados encontrados en la parte (*b*) usando las fórmulas (*3*) de este capítulo.

(*a*) Escogemos muestras sin remplazo para evitar los empates, que pueden ocurrir si, por ejemplo, el valor 12 apareciera en ambas muestras.

Hay $3 \cdot 2 = 6$ posibilidades de seleccionar las muestras, como se ve en la tabla 10-16. Se debe notar que podemos fácilmente usar los rangos 1, 2 y 3 en lugar de 7, 12 y 15. El valor U de la tabla 10-16 es el que se encontró para la muestra 1, pero si se hubiera usado U para la muestra 2, la distribución hubiese sido la misma.

Tabla 10-16

Muestra 1	Muestra 2	U
7	12 15	2
7	15 12	2
12	7 15	1
12	15 7	1
15	7 12	0
15	12 7	0

(*b*) La media y la varianza encontradas a partir de la tabla 10-15 están dadas por

$$\mu_U = \frac{2+2+1+1+0+0}{6} = 1$$

$$\sigma_U^2 = \frac{(2-1)^2+(2-1)^2+(1-1)^2+(1-1)^2+(0-1)^2+(0-1)^2}{6} = \frac{2}{3}$$

(*c*) Según las fórmulas (*3*),

$$\mu_U = \frac{N_1 N_2}{2} = \frac{(1)(2)}{2} = 1$$

$$\sigma_U^2 = \frac{N_1 N_2 (N_1 + N_2 + 1)}{12} = \frac{(1)(2)(1+2+1)}{12} = \frac{2}{3}$$

lo que muestra que concuerda con la parte (*a*).

10.11. (*a*) Encuentre la distribución muestral de U en el problema 10.9 y haga una gráfica.

(*b*) Obtenga la media y la varianza de U directamente a partir de los resultados de la parte (*a*).

(*c*) Verifique la parte (*b*) usando las fórmulas (*3*) de este capítulo.

(*a*) En este caso hay $5 \cdot 4 \cdot 3 \cdot 2 = 120$ posibilidades de escoger los valores de las dos muestras, y el método del problema 10.9 es demasiado laborioso. Para simplificar el procedimiento, concentrémonos en la muestra más pequeña (de tamaño $N_1 = 2$) y en la suma posible de los rangos R_1. La suma de los rangos para la muestra 1 es la *más pequeña* cuando la muestra consta de los dos números de rango más bajos (1, 2): entonces $R_1 = 1 + 2 = 3$. De igual manera, la suma de los rangos de la muestra 1 es la más grande cuando la muestra consta de los dos números de rango *más altos* (4, 5); entonces $R_1 = 4 + 5 = 9$. Así R_1 varía de 3 a 9.

La columna 1 de la tabla 10-17 enumera estos valores de R_1 (de 3 a 9) y la columna 2 muestra los valores correspondientes de la muestra 1, cuya suma es R_1. La columna 3 da la frecuencia (o número) de las muestras con suma R_1. Por ejemplo, hay $f = 2$ muestras con $R_1 = 5$. Dado que $N_1 = 2$ y $N_2 = 3$, tenemos

$$U = N_1 N_2 + \frac{N_1(N_1+1)}{2} - R_1 = (2)(3) + \frac{(2)(3)}{2} - R_1 = 9 - R_1$$

La probabilidad de que $U = R_1$ (es decir, $Pr\{U = R_1\}$) se muestra en la columna 5 de la tabla 10-17, y se obtuvo encontrando la frecuencia relativa. La frecuencia relativa se encontró dividiendo cada frecuencia f por la suma de todas las frecuencias, o 10. Por ejemplo, $Pr\{U = 5\} = \frac{2}{10} = 0.2$.

Tabla 10-17

R_1	Valores de la muestra 1	f	U	$\Pr\{U = R_1\}$
3	(1, 2)	1	6	0.1
4	(1, 3)	1	5	0.1
5	(1, 4), (2, 3)	2	4	0.2
6	(1, 5), (2, 4)	2	3	0.2
7	(2, 5), (3, 4)	2	2	0.2
8	(3, 5)	1	1	0.1
9	(4, 5)	1	0	0.1

(*b*) A partir de las columnas 3 y 4 de la tabla 10-17 tenemos

$$\mu_U = \bar{U} = \frac{\sum f\,U}{\sum f} = \frac{(1)(6)+(1)(5)+(2)(4)+(2)(3)+(2)(2)+(1)(1)+(1)(0)}{1+1+2+2+2+1+1} = 3$$

$$\sigma_U^2 = \frac{\sum f(U-\bar{U})^2}{\sum f}$$

$$= \frac{(1)(6-3)^2+(1)(5-3)^2+(2)(4-3)^2+(2)(3-3)^2+(2)(2-3)^2+(1)(1-3)^2+(1)(0-3)^2}{10}$$

$$= 3$$

Otro método.

$$\sigma_U^2 = \overline{U^2} - \bar{U}^2 = \frac{(1)(6)^2+(1)(5)^2+(2)(4)^2+(2)(3)^2+(2)(2)^2+(1)(1)^2+(1)(0)^2}{10} - (3)^2 = 3$$

(*c*) A partir de las fórmulas (*3*), usando $N_1 = 2$ y $N_2 = 3$, tenemos

$$\mu_U = \frac{N_1 N_2}{2} = \frac{(2)(3)}{2} = 3 \qquad \sigma_U^2 = \frac{N_1 N_2 (N_1 + N_2 + 1)}{12} = \frac{(2)(3)(6)}{12} = 3$$

10.12. Si N números de un conjunto se ordenan de 1 a N, demuestre que la suma de estos órdenes es $[N(N+1)]/2$.

Sea R la suma de los órdenes. Entonces tenemos

$$R = 1 + 2 + 3 + \cdots + (N-1) + N \tag{16}$$

$$R = N + (N-1) + (N-2) + \cdots + 2 + 1 \tag{17}$$

donde la suma en la ecuación (*17*) se obtiene escribiendo al revés la suma en (*16*). Agregando las ecuaciones (*16*) y (*17*) da

$$2R = (N+1) + (N+1) + (N+1) + \cdots + (N+1) + (N+1) = N(N+1)$$

dado que $(N + 1)$ ocurre N veces en la suma; así, $R = [N(N+1)]/2$. Esto también se puede obtener usando un resultado del álgebra elemental de las progresiones y series aritméticas.

10.13. Si R_1 y R_2 son las sumas respectivas de los órdenes de las muestras 1 y 2 en la prueba U, demuestre que $R_1 + R_2 = [N(N+1)]/2$.

Suponemos que no hay empates en los datos de la muestra. Entonces R_1 debe ser la suma de algunos de los rangos (números) en el conjunto 1, 2, 3, ..., N, mientras que R_2 debe ser la suma de todos los rangos en el conjunto. Esto es, $R_1 + R_2 = 1 + 2 + 3 + \ldots + N = [N(N+1)]/2$, por el problema 10.12.

LA PRUEBA *H* DE KRUSKAL-WALLIS

10.14. Una compañía quiere comprar una de cinco máquinas diferentes, *A*, *B*, *C*, *D* o *E*. En un experimento diseñado para determinar si hay un desempeño diferente entre las máquinas, cinco operadores experimentados trabajaron con cada máquina durante tiempos iguales. La tabla 10-18 muestra el número de unidades producidas por cada máquina. Pruebe la hipótesis al nivel de significancia de (*a*) 0.05, (*b*) 0.01 de que no hay diferencia entre las máquinas.

Tabla 10-18

A	68	72	77	42	53
B	72	53	63	53	48
C	60	82	64	75	72
D	48	61	57	64	50
E	64	65	70	68	53

Tabla 10-19

						Suma de los rangos
A	17.5	21	24	1	6.5	70
B	21	6.5	12	6.5	2.5	48.5
C	10	25	14	23	21	93
D	2.5	11	9	14	4	40.5
E	14	16	19	17.5	6.5	73

Como hay cinco muestras (*A*, *B*, *C*, *D*, y *E*), $k = 5$. Y dado que cada muestra consta de cinco valores, tenemos $N_1 = N_2 = N_3 = N_4 = N_5$, y $N = N_1 + N_2 + N_3 + N_4 + N_5 = 25$. Acomodando todos los valores de manera ascendente según su magnitud y asignando a los empates rangos apropiados, remplazamos la tabla 10-18 por la tabla 10-19, donde la columna de la derecha muestra la suma de los rangos. Vemos de la tabla 10-19 que $R_1 = 70$, $R_2 = 48.5$, $R_3 = 93$, $R_4 = 40.5$ y $R_5 = 73$. Así,

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{N_j} - 3(N+1)$$

$$= \frac{12}{(25)(26)} \left[\frac{(70)^2}{5} + \frac{(48.5)^2}{5} + \frac{(93)^2}{5} + \frac{(40.5)^2}{5} + \frac{(73)^2}{5} \right] - 3(26) = 6.44$$

Para $k - 1 = 4$ grados de libertad al nivel de significancia de 0.05, vemos del apéndice E que tenemos $\chi^2_{0.95} = 9.49$. Dado que $6.44 < 9.49$, no podemos rechazar la hipótesis de que no hay diferencia entre las máquinas al nivel de 0.05, y por lo tanto, tampoco podemos rechazarla al nivel de 0.01. En otras palabras, podemos aceptar la hipótesis (o abstenernos de hacer un juicio) de que no hay diferencia entre las máquinas a ambos niveles.

Observe que debido a que desarrollamos el problema haciendo un análisis de varianza (*véase* el problema 9.8) y llegamos a la misma conclusión.

10.15. Desarrolle el problema 10.14 si se hace una corrección para empates.

La tabla 10-20 muestra el número de empates correspondientes a cada observación empatada. Por ejemplo, 48 ocurre dos veces, por lo cual $T = 2$, y 53 ocurre cuatro veces por lo cual $T = 4$. Al calcular $T^3 - T$ para cada uno de estos valores de T y sumar, encontramos que $\sum(T^3 - T) = 6 + 60 + 24 + 6 + 24 = 120$, como se mostró en la tabla 10-20. Entonces, dado que $N = 25$, el factor de corrección es

$$1 - \frac{\sum(T^3 - T)}{N^3 - N} = 1 - \frac{120}{(25)^3 - 25} = 0.9923$$

Tabla 10-20

Observación	48	53	64	68	72	
Número de empates (T)	2	4	3	2	3	
$T^3 - T$	6	60	24	6	24	$\sum(T^3 - T) = 120$

y el valor corregido de H es

$$H_c = \frac{6.44}{0.9923} = 6.49$$

Esta corrección no es suficiente para cambiar la decisión tomada en el problema 10.14.

10.16. Se escogen al azar 3 muestras de una población. Acomodando los datos según rango obtenemos la tabla 10-21. Determine si hay alguna diferencia al nivel de significancia de (*a*) 0.05, (*b*) 0.01 entre las muestras.

Tabla 10-21

Muestra 1	7	4	6	10	
Muestra 2	11	9	12		
Muestra 3	5	1	3	8	2

Aquí $k = 3$, $N_1 = 4$, $N_2 = 3$, $N_3 = 5$, $N = N_1 + N_2 + N_3 = 12$, $R_1 = 7 + 4 + 6 + 10 = 27$, $R_2 = 11 + 9 + 12 = 32$, y $R_3 = 5 + 1 + 3 + 8 + 2 = 19$. Así

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{N_j} - 3(N+1) = \frac{12}{(12)(13)} \left[\frac{(27)^2}{4} + \frac{(32)^2}{3} + \frac{(19)^2}{5} \right] - 3(13) = 6.83$$

(*a*) Para $k - 1 = 3 - 1 = 2$ grados de libertad, $\chi^2_{0.95} = 5.99$. Así, dado que $6.83 > 5.99$, podemos concluir que hay una diferencia significativa entre las muestras al nivel de 0.05.

(*b*) Para 2 grados de libertad $\chi^2_{0.95} = 9.21$. Así, dado que $6.83 < 9.21$ no podemos concluir que hay una diferencia significativa entre las muestras al nivel de 0.01.

LA PRUEBA DE CORRIDAS PARA ALEATORIEDAD

10.17. En 30 lanzamientos de una moneda se encontró la siguiente secuencia de caras (H) y sellos (T):

H T T H T H H H T H H T T H T

H T H H T H T T H T H H T H T

(*a*) Determine el número de corridas, V.

(*b*) Pruebe al nivel de significancia de 0.05 si la secuencia es aleatoria.

(*a*) Usando una barra vertical para indicar una corrida, vemos de

H | T T | H | T | H H H | T | H H | T T | H | T |

H | T | H H | T | H | T T | H | T | H H | T | H |T |

que el número de corridas es $V = 22$.

(*b*) Hay $N_1 = 16$ caras y $N_2 = 14$ sellos en la muestra dada de lanzamientos, y de la parte (*a*), el número de corridas es $V = 22$. Por consiguiente, a partir de las fórmulas (*13*) de este capítulo, tenemos

$$\mu_V = \frac{2(16)(14)}{16+14} + 1 = 15.93 \quad ; \quad \sigma_V^2 = \frac{2(16)(14)[2(16)(14) - 16 - 14]}{(16+14)^2(16+14-1)} = 7.175$$

o $\sigma_V = 2.679$. El valor z correspondiente a $V = 22$ corridas es, por lo tanto,

$$z = \frac{V - \mu_V}{\sigma_V} = \frac{22 - 15.93}{2.679} = 2.27$$

Ahora, para una prueba de dos colas al nivel de significancia de 0.05, podemos aceptar la hipótesis H_0 de aleatoriedad si $-1.96 \leq z \leq 1.96$ y la rechazamos de otra manera (*véase* figura 10-4). Dado que el valor z

calculado es $2.27 > 1.96$, concluimos al nivel de 0.05 que los lanzamientos no son aleatorios. La prueba muestra que hay *demasiadas* corridas, que indican un *patrón cíclico* en los lanzamientos.

Si se usa la corrección por continuidad, el valor z anterior se remplaza por

$$z = \frac{(22 - 0.5) - 15.93}{2.679} = 2.08$$

y se obtiene la misma conclusión.

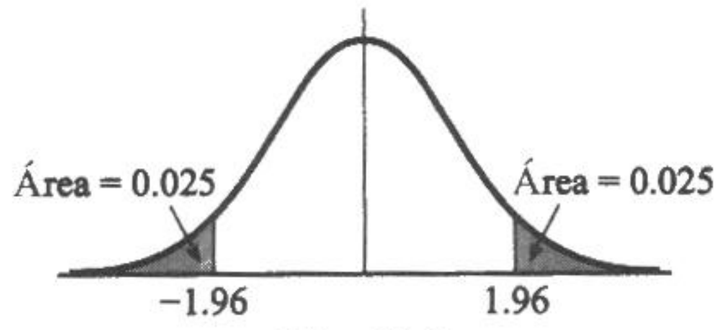


Fig. 10-4

10.18. Una muestra de 48 herramientas producidas por una máquina mostró la siguiente secuencia de buenas (G) y defectuosas (D):

G G G G G G D D G G G G G G G G

G G D D D D G G G G G G D G G G

G G G G G G D D G G G G G D G G

Pruebe la aleatoriedad de la secuencia al nivel de significancia de 0.05.

Las cantidades de letras D y G son $N_1 = 10$ y $N_2 = 38$, respectivamente, y el número de corridas es $V = 11$. Así, la media y la varianza están dadas por

$$\mu_V = \frac{2(10)(38)}{10 + 38} + 1 = 16.83 \quad ; \quad \sigma_V^2 = \frac{2(10)(38)[2(10)(38) - 10 - 38]}{(10 + 38)^2(10 + 38 - 1)} = 4.997$$

de manera que $\sigma_V = 2.235$.

Para una prueba de dos colas al nivel de significancia de 0.05, podemos aceptar la hipótesis H_0 de aleatoriedad si $-1.96 \leq z \leq 1.96$ (*véase* figura 10-4) y rechazarla de otra manera. Dado que el valor z correspondiente a $V = 11$ es

$$z = \frac{V - \mu_V}{\sigma_V} = \frac{11 - 16.83}{2.235} = -2.61$$

y $-2.61 < -1.96$, podemos rechazar H_0 al nivel de 0.05.

La prueba muestra que hay *muy pocas* corridas, lo que indica un amontonamiento (o apiñamiento) de tornillos defectuosos. En otras palabras, parece haber un *patrón de dispersión* en la producción de tornillos defectuosos. Se recomienda la observación posterior del proceso de producción.

10.19. (*a*) Forme todas las secuencias posibles que consten de 3 letras *a* y 2 *b*, y dé el número de corridas, V, correspondientes a cada secuencia.

(*b*) Obtenga la distribución muestral de V.

(*c*) Obtenga la distribución de probabilidad de V.

(*a*) El número de secuencias posibles que constan de 3 letras *a* y 2 *b* es

$$\binom{5}{2} = \frac{5!}{2!3!} = 10$$

Estas secuencias se muestran en la tabla 10-22, junto con el número de corridas correspondientes a cada secuencia.

Tabla 10-22

Secuencia	Corridas (V)
a a a b b	2
a a b a b	4
a a b b a	3
a b a b a	5
a b b a a	3
a b a a b	4
b b a a a	2
b a b a a	4
b a a a b	3
b a a b a	4

Tabla 10-23

V	f
2	2
3	3
4	4
5	1

(*b*) La distribución muestral de V está dada en la tabla 10-23 (obtenida de la tabla 10-22), donde V denota el número de corridas y f la frecuencia. Por ejemplo, la tabla 10-23 muestra que hay 1 cinco, 4 cuatros, etc.

(*c*) La distribución de probabilidad de V se obtiene de la tabla 10-23 dividiendo cada frecuencia por el total de la frecuencia $2 + 3 + 4 + 1 = 10$. Por ejemplo, $\Pr\{V = 5\} = \frac{1}{10} = 0.1$.

10.20. Encuentre (*a*) la media, (*b*) la varianza del número de corridas del problema 10.19 directamente de los resultados obtenidos allí.

(*a*) De la tabla 10-22 tenemos

$$\mu_V = \frac{2+4+3+5+3+4+2+4+3+4}{10} = \frac{17}{5}$$

Otro método.

De la tabla 10-22 el método de los datos agrupados da

$$\mu_V = \frac{\sum fV}{\sum f} = \frac{(2)(2)+(3)(3)+(4)(4)+(1)(5)}{2+3+4+1} = \frac{17}{5}$$

(*b*) Usando el método de los datos agrupados para calcular la varianza, de la tabla 10-23 tenemos

$$\sigma_V^2 = \frac{\sum f(V-\bar{V})^2}{\sum f} = \frac{1}{10}\left[(2)\left(2-\frac{17}{5}\right)^2 + (3)\left(3-\frac{17}{5}\right)^2 + (4)\left(4-\frac{17}{5}\right)^2 + (1)\left(5-\frac{17}{2}\right)^2\right] = \frac{21}{25}$$

Otro método.

Tal como en el capítulo 5, la varianza está dada por

$$\sigma_V^2 = \overline{V^2} - \bar{V}^2 = \frac{(2)(2)^2+(3)(3)^2+(4)(4)^2+(1)(5)^2}{10} - \left(\frac{17}{5}\right)^2 = \frac{21}{25}$$

10.21. Desarrolle el problema 10.20 usando las fórmulas (*13*) de este capítulo.

Dado que hay 3 letras *a* y 2 *b*, tenemos $N_1 = 3$ y $N_2 = 2$. Por lo tanto,

(*a*)
$$\mu_V = \frac{2N_1N_2}{N_1+N_2} + 1 = \frac{2(3)(2)}{3+2} + 1 = \frac{17}{5}$$

(*b*)
$$\sigma_V^2 = \frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1+N_2)^2(N_1+N_2-1)} = \frac{2(3)(2)[2(3)(2)-3-2]}{(3+2)^2(3+2-1)} = \frac{21}{25}$$

APLICACIONES ADICIONALES DE LA PRUEBA DE CORRIDAS

10.22. Con relación al problema 10.3 y suponiendo un nivel de significancia de 0.05, determine si los tiempos de vida de las muestras de las baterías producidas por el fabricante PQR son aleatorias.

La tabla 10-24 muestra los tiempos de vida de las baterías en orden ascendente de magnitud. Dado que hay 24 datos en la tabla, se obtiene la mediana a partir de los dos datos de la mitad, 253 y 262, como $(\frac{1}{2})(253 + 262) = 257.5$. Reescribiendo los datos de la tabla 10-23 usando una a si el dato está por encima de la mediana y una b, si está por debajo de la mediana, obtenemos la tabla 10-25, donde tenemos 12 letras a, 12 letras b y 15 corridas. Así, $N_1 = 12$, $N_2 = 12$, $V = 15$, y tenemos

$$\mu_V = \frac{2N_1N_2}{N_1 + N_2} + 1 = \frac{2(12)(12)}{12 + 12} + 1 = 13 \quad ; \quad \sigma_V^2 = \frac{2(12)(12)(264)}{(24)^2(23)} = 5.739$$

de manera que

$$z = \frac{V - \mu_V}{\sigma_V} = \frac{15 - 13}{2.396} = 0.835$$

Usando una prueba de dos colas al nivel de significancia de 0.05, podemos aceptar la hipótesis de aleatoriedad si $-1.96 \leq z \leq 1.96$. Dado que 0.835 cae en este rango, concluimos que la muestra es aleatoria.

Tabla 10-24

198	211	216	219	224	225	230	236
243	252	253	253	262	264	268	271
272	275	282	284	288	291	294	295

Tabla 10-25

a	*b*	*b*	*a*	*a*	*b*	*a*	*b*
b	*b*	*a*	*a*	*b*	*a*	*b*	*b*
a	*a*	*b*	*b*	*a*	*b*	*a*	*a*

10.23. Desarrolle el problema 10.5 usando la prueba de corridas para aleatoriedad.

El arreglo de los valores de ambas muestras aparece en la línea 1 de la tabla 10-8. Usando los símbolos a y b para los datos muestrales I y II, respectivamente, el arreglo se convierte en

$$b \quad b \quad b \quad b \quad b \quad b \quad b \quad b \quad a \quad a \quad a \quad a \quad a \quad b \quad b \quad a \quad a \quad a$$

Dado que hay cuatro corridas, tenemos $V = 4$, $N_1 = 8$ y $N_2 = 10$. Entonces

$$\mu_V = \frac{2N_1N_2}{N_1 + N_2} + 1 = \frac{2(8)(10)}{18} + 1 = 9.889$$

$$\sigma_V^2 = \frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1 + N_2)^2(N_1 + N_2 - 1)} = \frac{2(8)(10)(142)}{(18)^2(17)} = 4.125$$

de manera que

$$z = \frac{V - \mu_V}{\sigma_V} = \frac{4 - 9.889}{2.031} = -2.90$$

Si H_0 es la hipótesis de que no hay diferencia entre las aleaciones, también es la hipótesis de que la secuencia anterior es aleatoria. Aceptaremos esta hipótesis si $-1.96 \leq z \leq 1.96$ y la rechazaremos de otra manera. Dado que -2.90 está fuera de este intervalo, rechazamos H_0 y llegamos a la misma conclusión del problema 10.5.

Observe que si se hace una corrección por continuidad,

$$z = \frac{V - \mu_V}{\sigma_V} = \frac{(4 + 0.5) - 9.889}{2.031} = -2.65$$

y llegamos a la misma conclusión.

CORRELACIÓN DE RANGO

10.24. La tabla 10-26 muestra cómo 10 estudiantes, colocados en orden alfabético, se ordenaron según su desempeño tanto en el laboratorio como en la teoría de un curso de biología. Encuentre el coeficiente de correlación de rango.

Tabla 10-26

Laboratorio	8	3	9	2	7	10	4	6	1	5
Teoría	9	5	10	1	8	7	3	4	2	6

La diferencia en los rangos, D, en las secciones de laboratorio y teoría para cada estudiante aparece en la tabla 10-27, que también da D^2 y $\sum D^2$. Entonces,

$$r_S = 1 - \frac{6\sum D^2}{N(N^2-1)} = 1 - \frac{6(24)}{10(10^2-1)} = 0.8545$$

indica que hay una relación marcada entre el desempeño en teoría y laboratorio.

Tabla 10-27

Diferencia de rangos (D)	−1	−2	−1	1	−1	3	1	2	−1	−1	
D^2	1	4	1	1	1	9	1	4	1	1	$\sum D^2 = 24$

10.25. La tabla 10-28 muestra las estaturas de una muestra de 12 padres y sus hijos mayores. Encuentre el coeficiente de correlación rango.

Tabla 10-28

Estatura del padre (pulgadas)	65	63	67	64	68	62	70	66	68	67	69	71
Estatura del hijo (pulgadas)	68	66	68	65	69	66	68	65	71	67	68	70

Acomodados en orden ascendente de magnitud, las estaturas de los padres son

$$62 \quad 63 \quad 64 \quad 65 \quad 66 \quad 67 \quad 67 \quad 68 \quad 68 \quad 69 \quad 71 \tag{18}$$

Dado que los sitios sexto y séptimo en este orden representan la misma estatura (67 pulgadas), les asignamos un rango medio $(\frac{1}{2})(6 + 7) = 6.5$. De igual manera, los lugares octavo y noveno tienen asignado el rango $(\frac{1}{2})(8 + 9) = 8.5$. Así, a las estaturas de los padres se les asignan los rangos

$$1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6.5 \quad 6.5 \quad 8.5 \quad 8.5 \quad 10 \quad 11 \quad 12 \tag{19}$$

De igual manera, en orden ascendente de magnitud, las estaturas de los hijos son

$$65 \quad 65 \quad 66 \quad 66 \quad 67 \quad 68 \quad 68 \quad 68 \quad 68 \quad 69 \quad 70 \quad 71 \tag{20}$$

y dado que los lugares sexto, séptimo, octavo y noveno representan la misma estatura (68 pulgadas), les asignamos el rango medio $(\frac{1}{4})(6 + 7 + 8 + 9) = 7.5$. Así, a las estaturas de los hijos se les asignan los rangos

$$1.5 \quad 1.5 \quad 3.5 \quad 3.5 \quad 5 \quad 7.5 \quad 7.5 \quad 7.5 \quad 7.5 \quad 10 \quad 11 \quad 12 \tag{21}$$

Usando las correspondencias (*18*) y (*19*), y (*20*) y (*21*), podemos remplazar la tabla 10-28 por la 10-29. La tabla 10-30 muestra las diferencias en los rangos D y los cálculos de D^2 y $\sum D^2$, de donde

$$r_S = 1 - \frac{6\sum D^2}{N(N^2-1)} = 1 - \frac{6(72.50)}{12(12^2-1)} = 0.7465$$

El resultado concuerda bien con el coeficiente de correlación que se obtuvo por otros métodos (*véanse* problemas 8.26, 8.28, 8.30 y 8.32).

Tabla 10-29

Rango del padre	4	2	6.5	3	8.5	1	11	5	8.5	6.5	10	12
Rango del hijo	7.5	3.5	7.5	1.5	10	3.5	7.5	1.5	12	5	7.5	11

Tabla 10-30

D	−3.5	−1.5	−1.0	1.5	−1.5	−2.5	3.5	3.5	−3.5	1.5	2.5	1.0	
D^2	12.25	2.25	1.00	2.25	2.25	6.25	12.25	12.25	12.25	2.25	6.25	1.00	$\sum D^2 = 72.50$

Problemas suplementarios

LA PRUEBA DEL SIGNO

10.26. Una compañía afirma que si su producto se agrega al tanque de la gasolina, aumentará el kilometraje por galón. Para probar la afirmación, se seleccionaron 15 automóviles diferentes y se midió el kilometraje por galón con y sin el aditivo. Los resultados aparecen en la tabla 10-31. Suponiendo que las condiciones de conducción eran las mismas, determine si hay una diferencia significativa al nivel de (*a*) 0.05, (*b*) 0.01 debida al aditivo.

Tabla 10-31

Con aditivo	34.7	28.3	19.6	25.1	15.7	24.5	28.7	23.5	27.7	32.1	29.6	22.4	25.7	28.1	24.3
Sin aditivo	31.4	27.2	20.4	24.6	14.9	22.3	26.8	24.1	26.2	31.4	28.8	23.1	24.0	27.3	22.9

10.27. ¿Podemos concluir al nivel de significancia de 0.05 que el kilometraje por galón obtenido en el problema 10.26 es *mejor* con el aditivo que sin él?

10.28. En un aviso publicitario, un gimnasio anuncia que ha diseñado un programa especial que produce una pérdida de peso de al menos un 6% en un mes si se sigue detalladamente. Para probar la afirmación del gimnasio, 36 adultos emprendieron el programa. De ellos, 25 perdieron el peso deseado, 6 ganaron peso y el resto permaneció sin cambio. Determine al nivel de significancia de 0.05 si el programa es efectivo.

10.29. Un gerente de entrenamiento afirma que las ventas anuales de cierta compañía se incrementarán si el personal de ventas recibe un curso especial. Para probar esta afirmación, se da el curso a 24 personas. De esas 24 personas, se incrementan las ventas de 16, se disminuyen las de 6 y las de 2 permanecen iguales. Pruebe al nivel de significancia de 0.05 la hipótesis de que el curso aumenta las ventas de la compañía.

10.30. La fabrica de gaseosas MW lleva a cabo pruebas sensoriales en 27 localidades del país para determinar la preferencia relativa del público por dos marcas de gaseosas A y B. En ocho localidades se prefirió la marca A

sobre la *B*; en 17 localidades se prefirió la marca *B* sobre la *A*; y en las localidades restantes la preferencia fue indiferente. ¿Podemos concluir al nivel de significancia de 0.05 que la marca *B* se prefiere sobre la marca *A*?

10.31. La resistencia al rompimiento de una muestra aleatoria de 25 sogas hechas por un fabricante aparece en la tabla 10-32. Con base en esta muestra, pruebe al nivel de significancia de 0.05 la afirmación del fabricante de que la resistencia al rompimiento es (*a*) 25, (*b*) 30, (*c*) 35, (*d*) 40.

Tabla 10-32

41	28	35	38	23
37	32	24	46	30
25	36	22	41	37
43	27	34	27	36
42	33	28	31	24

10.32. Demuestre la manera de obtener los límites de confianza del 95% para los datos del problema 10.4.

10.33. Invente y resuelva un problema usando la prueba del signo.

LA PRUEBA *U* DE MANN-WHITNEY

10.34. Los profesores *A* y *B* dictan un curso básico de química en la universidad *XYZ*. En un examen final común, sus estudiantes recibieron las notas que aparecen en la tabla 10-33. Pruebe al nivel de significancia de 0.05 la hipótesis de que no hay diferencia entre las notas de los dos profesores.

Tabla 10-33

A	88	75	92	71	63	84	55	64	82	96				
B	72	65	84	53	76	80	51	60	57	85	94	87	73	61

10.35. Con relación al problema 10.34, ¿podemos concluir al nivel de significancia de 0.01 que las notas de los estudiantes en la clase de la mañana son peores que las de los de la tarde?

10.36. Un granjero desea determinar si hay diferencia en las producciones de dos variedades diferentes de trigo, I y II. La tabla 10-34 muestra la producción de trigo por unidad de área que usó cada variedad. ¿Puede concluir el granjero al nivel de significancia de (*a*) 0.05, (*b*) 0.01 que existe una diferencia?

Tabla 10-34

Trigo I	15.9	15.3	16.4	14.9	15.3	16.0	14.6	15.3	14.5	16.6	16.0
Trigo II	16.4	16.8	17.1	16.9	18.0	15.6	18.1	17.2	15.4		

10.37. ¿Puede el granjero del problema 10.36 concluir al nivel de 0.05 que el trigo II produce una cosecha más grande que el trigo I?

10.38. Una compañía quiere determinar si hay una diferencia entre dos marcas de gasolina, *A* y *B*. La tabla 10-35 muestra las distancias conducidas por galón para cada marca. ¿Podemos concluir al nivel de significancia de 0.05 que (*a*) hay una diferencia entre las marcas de gasolina, (*b*) que la marca *B* es mejor que la *A*?

Tabla 10-35

A	30.4	28.7	29.2	32.5	31.7	29.5	30.8	31.1	30.7	31.8
B	33.5	29.8	30.1	31.4	33.8	30.9	31.3	29.6	32.8	33.0

10.39. ¿Se puede usar la prueba U para determinar si hay diferencia entre las máquinas I y II de la tabla 10-1?

10.40. Invente y resuelva un problema usando la prueba U.

10.41. Encuentre U para los datos de la tabla 10-36, usando (*a*) el método de la fórmula, (*b*) el método de conteo.

Tabla 10-36

Muestra 1	15	25
Muestra 2	20	32

Tabla 10-37

Muestra 1	40	27	30	56
Muestra 2	10	35		

10.42. Desarrolle el problema 10.41 para los datos de la tabla 10-37.

10.43. Una población consta de los valores 2, 5, 9 y 12. Se toman dos muestras de esta población, siendo la primera uno de esos valores y la segunda, los otros tres.

(*a*) Obtenga la distribución muestral de U.

(*b*) Obtenga la media y la varianza de esta distribución, tanto directamente como a partir de una fórmula.

10.44. Demuestre que $U_1 + U_2 = N_1N_2$.

10.45. Demuestre que $R_1 + R_2 = [N(N + 1)]/2$ para el caso en que el número de empates sea (*a*) 1, (*b*) 2, (*c*) cualquiera.

10.46. Si $N_1 = 14$, $N_2 = 12$ y $R_1 = 105$, encuentre (*a*) R_2, (*b*) U_1, (*c*) U_2.

10.47. Si $N_1 = 10$, $N_2 = 16$ y $U_2 = 60$, encuentre (*a*) R_1, (*b*) R_2, (*c*) U_1.

10.48. ¿Cuál es el número más grande de los valores N_1, N_2, R_1, R_2, U_1 y U_2 que se puede determinar a partir de los restantes? Demuestre su respuesta.

LA PRUEBA *H* DE KRUSKAL-WALLIS

10.49. Se lleva a cabo un experimento para determinar la producción de cinco variedades de trigo *A*, *B*, *C*, *D* y *E*. A cada variedad se le asignan cuatro parcelas. La producción (en toneladas por hectárea) se muestra en la tabla 10-38. Suponiendo que las parcelas tienen fertilidad similar y que se les asignan variedades al azar, determine si hay diferencia significativa entre la producción al nivel de (*a*) 0.05, (*b*) 0.01.

Tabla 10-38

A	20	12	15	19
B	17	14	12	15
C	23	16	18	14
D	15	17	20	12
E	21	14	17	18

Tabla 10-39

A	33	38	36	40	31	35
B	32	40	42	38	30	34
C	31	37	35	33	34	30
D	27	33	32	29	31	28

10.50. Una compañía quiere probar cuatro tipos diferentes de llantas: *A*, *B*, *C* y *D*. Los tiempos de vida de las llantas, como se determina a partir de sus fibras, se dan (en miles de kilómetros) en la tabla 10-39. Cada tipo se probó en seis automóviles similares asignados al azar a las llantas. Determine si hay una diferencia significativa entre las llantas a los niveles de (*a*) 0.05, (*b*) 0.01.

10.51. Una profesora quiere probar tres métodos diferentes de enseñanza: I, II y III. Para hacerlo, la profesora selecciona al azar tres grupos de cinco estudiantes cada uno, y dicta sus clases en cada grupo usando métodos diferentes. Se da a todos los estudiantes el mismo examen y se obtienen las notas de la tabla 10-40. Determine al nivel de significancia de (*a*) 0.05, (*b*) 0.01 si hay una diferencia entre los métodos de enseñanza.

Tabla 10-40

Método I	78	62	71	58	73
Método II	76	85	77	90	87
Método III	74	79	60	75	80

10.52. Durante un semestre un estudiante recibió en varias materias las calificaciones que aparecen en la tabla 10-41. Pruebe a los niveles de significancia de (*a*) 0.05, (*b*) 0.01 si hay una diferencia entre las notas de esta persona.

Tabla 10-41

Matemáticas	72	80	83	75	
Biología	81	74	77		
Inglés	88	82	90	87	80
Economía	74	71	77	70	

10.53. Usando la prueba *H* desarrolle (*a*) el problema 9.14, (*b*) el problema 9.23, (*c*) el problema 9.24.

10.54. Usando la prueba *H* desarrolle (*a*) el problema 9.25, (*b*) el problema 9.26, (*c*) el problema 9.27.

LA PRUEBA DE CORRIDAS PARA ALEATORIEDAD

10.55. Determine el número de corridas, *V*, para cada una de las siguientes secuencias:

(*a*) **A B A B B A A A B B A B**

(*b*) **H H T H H H T T T T H H T H H T H T**

10.56. Se muestrean 25 individuos que dan su opinión acerca de si les gusta o no cierto producto (indicado como Y o N, respectivamente). Los datos resultantes se muestran en la siguiente secuencia:

Y Y N N N N Y Y Y N Y N N Y N N N N N Y Y Y Y N N

(*a*) Determine el número de corridas, *V*.

(*b*) Pruebe al nivel de significancia de 0.05 si las respuestas son aleatorias.

10.57. Use la prueba de corridas (*10*) y (*11*) de este capítulo, y plantee sus conclusiones sobre aleatoriedad.

10.58. (*a*) Forme todas las secuencias posibles que consten de dos letras *a* y una letra *b*, y dé el número de corridas, *V*, correspondiente a cada secuencia.

(*b*) Obtenga la distribución muestral de *V*.

(*c*) Obtenga la distribución de probabilidad de *V*.

10.59. En el problema 10.58, encuentre la media y la varianza de *V* (*a*) directamente a partir de la distribución muestral, (*b*) por fórmula.

10.60. Desarrolle los problemas 10.58 y 10.59 para los casos en los cuales hay (*a*) dos letras *a* y dos *b*, (*b*) una letra *a* y tres *b*, (*c*) una letra *a* y cuatro *b*.

10.61. Desarrolle los problemas 10.58 y 10.59 para los casos en los cuales hay (*a*) dos letras *a* y cuatro *b*, (*b*) tres letras *a* y tres *b*.

APLICACIONES ADICIONALES DE LA PRUEBA DE CORRIDAS

10.62. Suponiendo un nivel de significancia de 0.05, determine si la muestra de las 40 notas de la tabla 10-5 es aleatoria.

10.63. En la tabla 10-42 se muestran los precios a los que cerraron las acciones en 25 días consecutivos. Determine al nivel de significancia de 0.05 si los precios son aleatorios.

Tabla 10-42

10.375	11.125	10.875	10.625	11.500
11.625	11.250	11.375	10.750	11.000
10.875	10.750	11.500	11.250	12.125
11.875	11.375	11.875	11.125	11.750
11.375	12.125	11.750	11.500	12.250

10.64. Los primeros dígitos de $\sqrt{2}$ son 1.41421 35623 73095 0488..., ¿A qué conclusiones llega con relación a la aleatoriedad de los dígitos?

10.65. ¿Qué conclusiones puede usted obtener con relación a la aleatoriedad de los siguientes dígitos?

(*a*) $\sqrt{3} = 1.73205\ 08075\ 68877\ 2935\cdots$

(*b*) $\pi = 3.14159\ 26535\ 89793\ 2643\cdots$

10.66. Desarrolle el problema 10.30 usando la prueba de corridas para aleatoriedad.

10.67. Desarrolle el problema 10.32 usando la prueba de corridas para aleatoriedad.

10.68. Desarrolle el problema 10.34 usando la prueba de corridas para aleatoriedad.

CORRELACIÓN DE RANGO

10.69. En un concurso, se pidió a dos jueces que colocaran ocho candidatos (numerados de 1 a 8) en orden de preferencia. Las preferencias de los jueces aparecen en la tabla 10-43.

(*a*) Encuentre el coeficiente de correlación de rango.

(*b*) Decida qué tanto concuerdan los jueces en sus elecciones.

Tabla 10-43

Primer juez	5	2	8	1	4	6	3	7
Segundo juez	4	5	7	3	2	8	1	6

10.70. El coeficiente de correlación de rango se deriva usando los datos ordenados en la fórmula de producto-momento del capítulo 8. Ilustre esto usando ambos métodos para trabajar el problema.

10.71. ¿Se puede encontrar el coeficiente de correlación de rango para datos agrupados? Explique esto e ilustre su respuesta con un ejemplo.

Respuestas a los problemas suplementarios

10.26. Hay una diferencia significativa al nivel de 0.05, pero no al nivel de 0.01.

10.27. Sí.

10.28. Al nivel de 0.05 el programa es efectivo.

10.29. Podemos rechazar la hipótesis de las ventas incrementadas al nivel de 0.05.

10.30. No.

10.31. (*a*) Rechazar. (*b*) Aceptar. (*c*) Aceptar. (*d*) Rechazar.

10.34. No hay diferencia al nivel de 0.05.

10.35. No.

10.36. (*a*) Sí. (*b*) Sí.

10.37. Sí.

10.38. (*a*) Sí. (*b*) Sí.

10.41. 3.

10.42. 6.

10.49. No hay diferencia significativa a ningún nivel.

10.50. La diferencia es significativa al nivel de 0.05, pero no al nivel de 0.01.

10.51. La diferencia es significativa al nivel de 0.05, pero no al nivel de 0.01.

10.52. Hay diferencia significativa en las notas a ambos niveles.

10.55. (*a*) 8. (*b*) 10.

10.56. (*a*) 10. (*b*) Las respuestas son al azar al nivel de 0.05.

10.62. La muestra no es aleatoria al nivel de 0.05. Hay *demasiadas* corridas que indican un patrón cíclico.

10.63. La muestra no es aleatoria al nivel de 0.05. Hay *muy pocas* corridas que indican un patrón disperso.

10.64. Los dígitos son aleatorios al nivel de 0.05.

10.65. (*a*) Los dígitos son aleatorios al nivel de 0.05. (*b*) Los dígitos son aleatorios al nivel de 0.05.

10.69. (*a*) 0.67. (*b*) Los jueces no concuerdan estrechamente en sus elecciones.

APÉNDICE A

Temas matemáticos

SUMAS ESPECIALES

Las siguientes son algunas sumas de series que surgen en la práctica. Por definición, $0! = 1$. Cuando la serie es infinita, se indica el rango de convergencia.

1. $\displaystyle\sum_{j=1}^{m} j = 1 + 2 + 3 + \cdots + m = \frac{m(m+1)}{2}$

2. $\displaystyle\sum_{j=1}^{m} j^2 = 1^2 + 2^2 + 3^2 + \cdots + m^2 = \frac{m(m+1)(2m+1)}{6}$

3. $\displaystyle e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \sum_{j=0}^{\infty} \frac{x^j}{j!}$ para todo x

4. $\displaystyle \operatorname{sen} x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots = \sum_{j=0}^{\infty} \frac{(-1)^j x^{2j+1}}{(2j+1)!}$ para todo x

5. $\displaystyle \cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots = \sum_{j=0}^{\infty} \frac{(-1)^j x^{2j}}{(2j)!}$ para todo x

6. $\displaystyle \frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots = \sum_{j=0}^{\infty} x^j$ $|x| < 1$

7. $\displaystyle \ln(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \cdots = -\sum_{j=1}^{\infty} \frac{x^j}{j}$ $-1 \leqq x < 1$

FÓRMULAS DE EULER

8. $e^{i\theta} = \cos\theta + i \operatorname{sen}\theta, \qquad e^{-i\theta} = \cos\theta - i \operatorname{sen}\theta$

9. $\displaystyle \cos\theta = \frac{e^{i\theta} + e^{-i\theta}}{2}, \qquad \operatorname{sen}\theta = \frac{e^{i\theta} - e^{-i\theta}}{2i}$

LA FUNCIÓN GAMMA

La *función gamma,* denotada por $\Gamma(n)$, se define como

$$\Gamma(n) = \int_0^{\infty} t^{n-1} e^{-t}\, dt \qquad n > 0 \tag{1}$$

Una *fórmula de recurrencia* está dada por

$$\Gamma(n+1) = n\Gamma(n) \tag{2}$$

donde $\Gamma(1) = 1$. Se puede obtener una extensión de la función gamma para $n < 0$ usando (2).

Si n es un entero positivo, entonces

$$\Gamma(n+1) = n! \tag{3}$$

Por esta razón, $\Gamma(n)$ se denomina algunas veces la función factorial. Una propiedad importante de la función gamma es

$$\Gamma(p)\Gamma(1-p) = \frac{\pi}{\operatorname{sen} p\pi} \tag{4}$$

Para $p = \frac{1}{2}$, (4) da

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi} \tag{5}$$

Para valores grandes de n tenemos la *fórmula asintóntica de Stirling*:

$$\Gamma(n+1) \sim \sqrt{2\pi n}\, n^n e^{-n} \tag{6}$$

donde el signo $\sim$ indica que la razón de los dos lados se aproxima a 1 cuando $n \to \infty$. En particular, si n es un entero positivo grande, una buena aproximación para $n!$ está dada por:

$$n! \sim \sqrt{2\pi n}\, n^n e^{-n} \tag{7}$$

LA FUNCIÓN BETA

La *función beta*, denotada por $B(m, n)$, se define como

$$B(m,n) = \int_0^1 u^{m-1}(1-u)^{n-1}\,du \qquad m > 0, n > 0 \tag{8}$$

Ella se relaciona con la función gamma por

$$B(m,n) = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)} \tag{9}$$

INTEGRALES ESPECIALES

Las siguientes son algunas integrales que surgen en probabilidad y estadística.

10. $\displaystyle\int_0^\infty e^{-ax^2}\,dx = \frac{1}{2}\sqrt{\frac{\pi}{a}} \quad ; \quad a > 0$

11. $\displaystyle\int_0^\infty x^m e^{-ax^2}\,dx = \frac{\Gamma\left(\frac{m+1}{2}\right)}{2a^{(m+1)/2}} \quad ; \quad a > 0,\ m > -1$

12. $\displaystyle\int_0^\infty e^{-ax^2}\cos bx\,dx = \frac{1}{2}\sqrt{\frac{\pi}{a}}\, e^{-b^2/4a} \quad ; \quad a > 0$

13. $\displaystyle\int_0^\infty e^{-ax}\cos bx\ dx = \frac{a}{a^2+b^2} \quad ; \quad a>0$

14. $\displaystyle\int_0^\infty e^{-ax}\,\text{sen}\,bx\ dx = \frac{b}{a^2+b^2} \quad ; \quad a>0$

15. $\displaystyle\int_0^\infty x^{p-1}e^{-ax}\,dx = \frac{\Gamma(p)}{a^p} \quad ; \quad a>0,\ p>0$

16. $\displaystyle\int_{-\infty}^\infty e^{-(ax^2+bx+c)}\,dx = \sqrt{\frac{\pi}{a}}\,e^{(b^2-4ac)/4a} \quad ; \quad a>0$

17. $\displaystyle\int_0^\infty e^{-(ax^2+bx+c)}\,dx = \frac{1}{2}\sqrt{\frac{\pi}{a}}\,e^{(b^2-4ac)/4a}\,\text{fcer}\left(\frac{b}{2\sqrt{a}}\right) \quad ; \quad a>0$

donde

$$\text{fcer}(u) = 1 - \ \text{fer}(u) = 1 - \frac{2}{\sqrt{\pi}}\int_0^u e^{-x^2}\,dx = \frac{2}{\sqrt{\pi}}\int_u^\infty e^{-x^2}\,dx$$

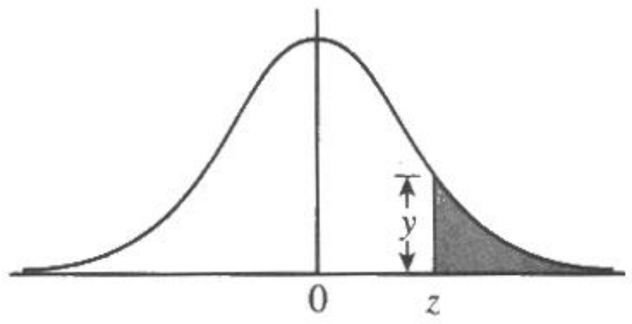
se conoce como la *función complementaria de error.*

18. $\displaystyle\int_0^\infty \frac{\cos\omega x}{x^2+a^2}\,dx = \frac{\pi}{2a}\,e^{-a\omega} \quad ; \quad a>0,\ \omega>0$

19. $\displaystyle\int_0^{\pi/2} \text{sen}^{2m-1}\theta\cos^{2n-1}\theta\ d\theta = \frac{\Gamma(m)\Gamma(n)}{2\Gamma(m+n)} \quad ; \quad m>0,\ n>0$

APÉNDICE B

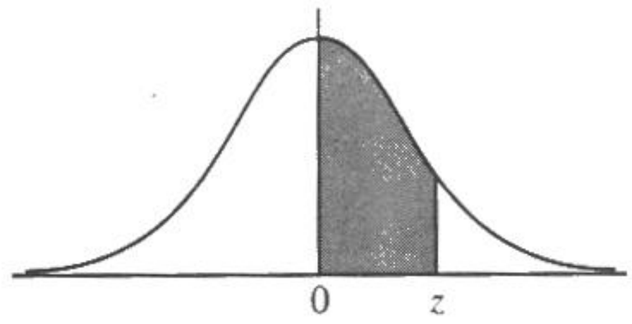
Ordenadas *y* de la curva normal estándar en *z*



z	0	1	2	3	4	5	6	7	8	9
0.0	.3989	.3989	.3989	.3988	.3986	.3984	.3982	.3980	.3977	.3973
0.1	.3970	.3965	.3961	.3956	.3951	.3945	.3939	.3932	.3925	.3918
0.2	.3910	.3902	.3894	.3885	.3876	.3867	.3857	.3847	.3836	.3825
0.3	.3814	.3802	.3790	.3778	.3765	.3752	.3739	.3725	.3712	.3697
0.4	.3683	.3668	.3653	.3637	.3621	.3605	.3589	.3572	.3555	.3538
0.5	.3521	.3503	.3485	.3467	.3448	.3429	.3410	.3391	.3372	.3352
0.6	.3332	.3312	.3292	.3271	.3251	.3230	.3209	.3187	.3166	.3144
0.7	.3123	.3101	.3079	.3056	.3034	.3011	.2989	.2966	.2943	.2920
0.8	.2897	.2874	.2850	.2827	.2803	.2780	.2756	.2732	.2709	.2685
0.9	.2661	.2637	.2613	.2589	.2565	.2541	.2516	.2492	.2468	.2444
1.0	.2420	.2396	.2371	.2347	.2323	.2299	.2275	.2251	.2227	.2203
1.1	.2179	.2155	.2131	.2107	.2083	.2059	.2036	.2012	.1989	.1965
1.2	.1942	.1919	.1895	.1872	.1849	.1826	.1804	.1781	.1758	.1736
1.3	.1714	.1691	.1669	.1647	.1626	.1604	.1582	.1561	.1539	.1518
1.4	.1497	.1476	.1456	.1435	.1415	.1394	.1374	.1354	.1334	.1315
1.5	.1295	.1276	.1257	.1238	.1219	.1200	.1182	.1163	.1145	.1127
1.6	.1109	.1092	.1074	.1057	.1040	.1023	.1006	.0989	.0973	.0957
1.7	.0940	.0925	.0909	.0893	.0878	.0863	.0848	.0833	.0818	.0804
1.8	.0790	.0775	.0761	.0748	.0734	.0721	.0707	.0694	.0681	.0669
1.9	.0656	.0644	.0632	.0620	.0608	.0596	.0584	.0573	.0562	.0551
2.0	.0540	.0529	.0519	.0508	.0498	.0488	.0478	.0468	.0459	.0449
2.1	.0440	.0431	.0422	.0413	.0404	.0396	.0387	.0379	.0371	.0363
2.2	.0355	.0347	.0339	.0332	.0325	.0317	.0310	.0303	.0297	.0290
2.3	.0283	.0277	.0270	.0264	.0258	.0252	.0246	.0241	.0235	.0229
2.4	.0224	.0219	.0213	.0208	.0203	.0198	.0194	.0189	.0184	.0180
2.5	.0175	.0171	.0167	.0163	.0158	.0154	.0151	.0147	.0143	.0139
2.6	.0136	.0132	.0129	.0126	.0122	.0119	.0116	.0113	.0110	.0107
2.7	.0104	.0101	.0099	.0096	.0093	.0091	.0088	.0086	.0084	.0081
2.8	.0079	.0077	.0075	.0073	.0071	.0069	.0067	.0065	.0063	.0061
2.9	.0060	.0058	.0056	.0055	.0053	.0051	.0050	.0048	.0047	.0046
3.0	.0044	.0043	.0042	.0040	.0039	.0038	.0037	.0036	.0035	.0034
3.1	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026	.0025	.0025
3.2	.0024	.0023	.0022	.0022	.0021	.0020	.0020	.0019	.0018	.0018
3.3	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014	.0013	.0013
3.4	.0012	.0012	.0012	.0011	.0011	.0010	.0010	.0010	.0009	.0009
3.5	.0009	.0008	.0008	.0008	.0008	.0007	.0007	.0007	.0007	.0006
3.6	.0006	.0006	.0006	.0005	.0005	.0005	.0005	.0005	.0005	.0004
3.7	.0004	.0004	.0004	.0004	.0004	.0004	.0003	.0003	.0003	.0003
3.8	.0003	.0003	.0003	.0003	.0003	.0002	.0002	.0002	.0002	.0002
3.9	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0001	.0001

APÉNDICE C

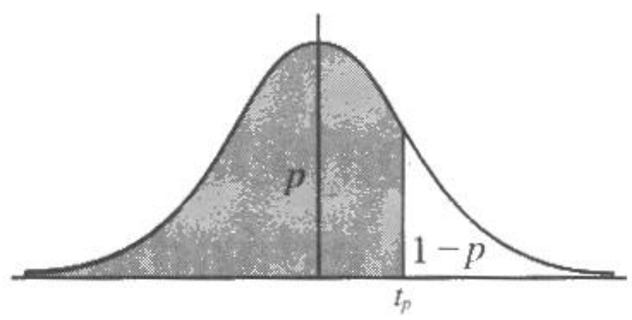
Áreas bajo la curva normal estándar de 0 a z



z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0754
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2258	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2996	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000

APÉNDICE D

Valores percentiles t_p para la distribución t de Student con ν grados de libertad

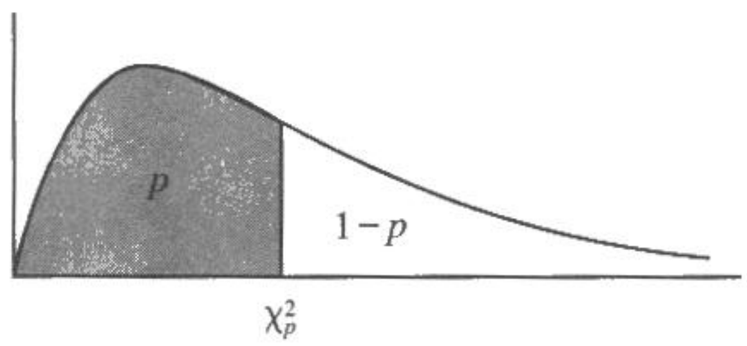


ν	$t_{.55}$	$t_{.60}$	$t_{.70}$	$t_{.75}$	$t_{.80}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$
1	.158	.325	.727	1.000	1.376	3.08	6.31	12.71	31.82	63.66
2	.142	.289	.617	.816	1.061	1.89	2.92	4.30	6.96	9.92
3	.137	.277	.584	.765	.978	1.64	2.35	3.18	4.54	5.84
4	.134	.271	.569	.741	.941	1.53	2.13	2.78	3.75	4.60
5	.132	.267	.559	.727	.920	1.48	2.02	2.57	3.36	4.03
6	.131	.265	.553	.718	.906	1.44	1.94	2.45	3.14	3.71
7	.130	.263	.549	.711	.896	1.42	1.90	2.36	3.00	3.50
8	.130	.262	.546	.706	.889	1.40	1.86	2.31	2.90	3.36
9	.129	.261	.543	.703	.883	1.38	1.83	2.26	2.82	3.25
10	.129	.260	.542	.700	.879	1.37	1.81	2.23	2.76	3.17
11	.129	.260	.540	.697	.876	1.36	1.80	2.20	2.72	3.11
12	.128	.259	.539	.695	.873	1.36	1.78	2.18	2.68	3.06
13	.128	.259	.538	.694	.870	1.35	1.77	2.16	2.65	3.01
14	.128	.258	.537	.692	.868	1.34	1.76	2.14	2.62	2.98
15	.128	.258	.536	.691	.866	1.34	1.75	2.13	2.60	2.95
16	.128	.258	.535	.690	.865	1.34	1.75	2.12	2.58	2.92
17	.128	.257	.534	.689	.863	1.33	1.74	2.11	2.57	2.90
18	.127	.257	.534	.688	.862	1.33	1.73	2.10	2.55	2.88
19	.127	.257	.533	.688	.861	1.33	1.73	2.09	2.54	2.86
20	.127	.257	.533	.687	.860	1.32	1.72	2.09	2.53	2.84
21	.127	.257	.532	.686	.859	1.32	1.72	2.08	2.52	2.83
22	.127	.256	.532	.686	.858	1.32	1.72	2.07	2.51	2.82
23	.127	.256	.532	.685	.858	1.32	1.71	2.07	2.50	2.81
24	.127	.256	.531	.685	.857	1.32	1.71	2.06	2.49	2.80
25	.127	.256	.531	.684	.856	1.32	1.71	2.06	2.48	2.79
26	.127	.256	.531	.684	.856	1.32	1.71	2.06	2.48	2.78
27	.127	.256	.531	.684	.855	1.31	1.70	2.05	2.47	2.77
28	.127	.256	.530	.683	.855	1.31	1.70	2.05	2.47	2.76
29	.127	.256	.530	.683	.854	1.31	1.70	2.04	2.46	2.76
30	.127	.256	.530	.683	.854	1.31	1.70	2.04	2.46	2.75
40	.126	.255	.529	.681	.851	1.30	1.68	2.02	2.42	2.70
60	.126	.254	.527	.679	.848	1.30	1.67	2.00	2.39	2.66
120	.126	.254	.526	.677	.845	1.29	1.66	1.98	2.36	2.62
∞	.126	.253	.524	.674	.842	1.28	1.645	1.96	2.33	2.58

Fuente: R. A. Fisher y F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, publicado por Longman Group, Ltd., Londres (previamente publicado por Oliver y Boyd, Edinburgo), y con autorización de los autores y editores.

APÉNDICE E

Valores percentiles χ_p^2 para la distribución chi cuadrado con ν grados de libertad

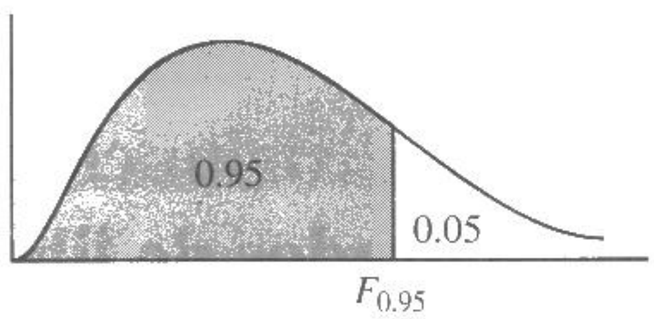


ν	$\chi^2_{.005}$	$\chi^2_{.01}$	$\chi^2_{.025}$	$\chi^2_{.05}$	$\chi^2_{.10}$	$\chi^2_{.25}$	$\chi^2_{.50}$	$\chi^2_{.75}$	$\chi^2_{.90}$	$\chi^2_{.95}$	$\chi^2_{.975}$	$\chi^2_{.99}$	$\chi^2_{.995}$	$\chi^2_{.999}$
1	.0000	.0002	.0010	.0039	.0158	.102	.455	1.32	2.71	3.84	5.02	6.63	7.88	10.8
2	.0100	.0201	.0506	.103	.211	.575	1.39	2.77	4.61	5.99	7.38	9.21	10.6	13.8
3	.0717	.115	.216	.352	.584	1.21	2.37	4.11	6.25	7.81	9.35	11.3	12.8	16.3
4	.207	.297	.484	.711	1.06	1.92	3.36	5.39	7.78	9.49	11.1	13.3	14.9	18.5
5	.412	.554	.831	1.15	1.61	2.67	4.35	6.63	9.24	11.1	12.8	15.1	16.7	20.5
6	.676	.872	1.24	1.64	2.20	3.45	5.35	7.84	10.6	12.6	14.4	16.8	18.5	22.5
7	.989	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.0	14.1	16.0	18.5	20.3	24.3
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.2	13.4	15.5	17.5	20.1	22.0	26.1
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.4	14.7	16.9	19.0	21.7	23.6	27.9
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.5	16.0	18.3	20.5	23.2	25.2	29.6
11	2.60	3.05	3.82	4.57	5.58	7.58	10.3	13.7	17.3	19.7	21.9	24.7	26.8	31.3
12	3.07	3.57	4.40	5.23	6.30	8.44	11.3	14.8	18.5	21.0	23.3	26.2	28.3	32.9
13	3.57	4.11	5.01	5.89	7.04	9.30	12.3	16.0	19.8	22.4	24.7	27.7	29.8	34.5
14	4.07	4.66	5.63	6.57	7.79	10.2	13.3	17.1	21.1	23.7	26.1	29.1	31.3	36.1
15	4.60	5.23	6.26	7.26	8.55	11.0	14.3	18.2	22.3	25.0	27.5	30.6	32.8	37.7
16	5.14	5.81	6.91	7.96	9.31	11.9	15.3	19.4	23.5	26.3	28.8	32.0	34.3	39.3
17	5.70	6.41	7.56	8.67	10.1	12.8	16.3	20.5	24.8	27.6	30.2	33.4	35.7	40.8
18	6.26	7.01	8.23	9.39	10.9	13.7	17.3	21.6	26.0	28.9	31.5	34.8	37.2	42.3
19	6.84	7.63	8.91	10.1	11.7	14.6	18.3	22.7	27.2	30.1	32.9	36.2	38.6	43.8
20	7.43	8.26	9.59	10.9	12.4	15.5	19.3	23.8	28.4	31.4	34.2	37.6	40.0	45.3
21	8.03	8.90	10.3	11.6	13.2	16.3	20.3	24.9	29.6	32.7	35.5	38.9	41.4	46.8
22	8.64	9.54	11.0	12.3	14.0	17.2	21.3	26.0	30.8	33.9	36.8	40.3	42.8	48.3
23	9.26	10.2	11.7	13.1	14.8	18.1	22.3	27.1	32.0	35.2	38.1	41.6	44.2	49.7
24	9.89	10.9	12.4	13.8	15.7	19.0	23.3	28.2	33.2	36.4	39.4	43.0	45.6	51.2
25	10.5	11.5	13.1	14.6	16.5	19.9	24.3	29.3	34.4	37.7	40.6	44.3	46.9	52.6
26	11.2	12.2	13.8	15.4	17.3	20.8	25.3	30.4	35.6	38.9	41.9	45.6	48.3	54.1
27	11.8	12.9	14.6	16.2	18.1	21.7	26.3	31.5	36.7	40.1	43.2	47.0	49.6	55.5
28	12.5	13.6	15.3	16.9	18.9	22.7	27.3	32.6	37.9	41.3	44.5	48.3	51.0	56.9
29	13.1	14.3	16.0	17.7	19.8	23.6	28.3	33.7	39.1	42.6	45.7	49.6	52.3	58.3
30	13.8	15.0	16.8	18.5	20.6	24.5	29.3	34.8	40.3	43.8	47.0	50.9	53.7	59.7
40	20.7	22.2	24.4	26.5	29.1	33.7	39.3	45.6	51.8	55.8	59.3	63.7	66.8	73.4
50	28.0	29.7	32.4	34.8	37.7	42.9	49.3	56.3	63.2	67.5	71.4	76.2	79.5	86.7
60	35.5	37.5	40.5	43.2	46.5	52.3	59.3	67.0	74.4	79.1	83.3	88.4	92.0	99.6
70	43.3	45.4	48.8	51.7	55.3	61.7	69.3	77.6	85.5	90.5	95.0	100	104	112
80	51.2	53.5	57.2	60.4	64.3	71.1	79.3	88.1	96.6	102	107	112	116	125
90	59.2	61.8	65.6	69.1	73.3	80.6	89.3	98.6	108	113	118	124	128	137
100	67.3	70.1	74.2	77.9	82.4	90.1	99.3	109	118	124	130	136	140	149

Fuente: E. S. Pearson y H. O. Hartley, *Biometrika Tables for Statisticians*, Vol. 1 (1966), tabla 8, páginas 137 y 138, con autorización.

APÉNDICE F

Percentiles 95 (niveles de 0.05), $F_{0.95}$, para la distribución F

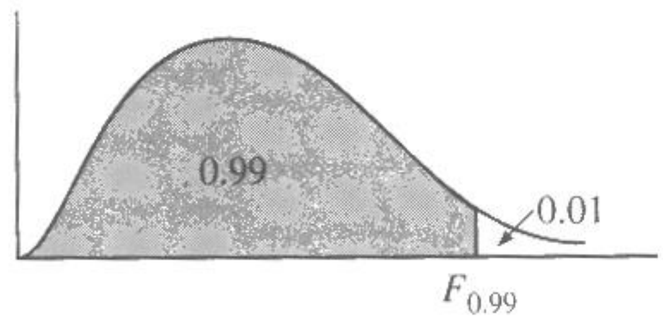


ν_1 grados de libertad en el numerador
ν_2 grados de libertad en el denominador

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Fuente: E. S. Pearson y H. O. Hartley, *Biometrika Tables for Statisticians*, Vol. 2 (1972), tabla 5, página 178, con autorización.

Percentiles 99 (niveles de 0.01), $F_{0.99}$, para la distribución F



ν_1 grados de libertad en el numerador
ν_2 grados de libertad en el denominador

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	5000	5403	5625	5764	5859	5928	5981	6023	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.9	13.8	13.7	13.7	13.6	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.82	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Fuente: E. S. Pearson y H. O. Hartley, *Biometrika Tables for Statisticians*, Vol. 2 (1972), tabla 5, página 180 , con autorización.

APÉNDICE G

Valores de $e^{-\lambda}$

$(0 < \lambda < 1)$

λ	0	1	2	3	4	5	6	7	8	9
0.0	1.0000	.9900	.9802	.9704	.9608	.9512	.9418	.9324	.9231	.9139
0.1	.9048	.8958	.8869	.8781	.8694	.8607	.8521	.8437	.8353	.8270
0.2	.8187	.8106	.8025	.7945	.7866	.7788	.7711	.7634	.7558	.7483
0.3	.7408	.7334	.7261	.7189	.7118	.7047	.6977	.6907	.6839	.6771
0.4	.6703	.6636	.6570	.6505	.6440	.6376	.6313	.6250	.6188	.6126
0.5	.6065	.6005	.5945	.5886	.5827	.5770	.5712	.5655	.5599	.5543
0.6	.5488	.5434	.5379	.5326	.5273	.5220	.5169	.5117	.5066	.5016
0.7	.4966	.4916	.4868	.4819	.4771	.4724	.4677	.4630	.4584	.4538
0.8	.4493	.4449	.4404	.4360	.4317	.4274	.4232	.4190	.4148	.4107
0.9	.4066	.4025	.3985	.3946	.3906	.3867	.3829	.3791	.3753	.3716

$(\lambda = 1, 2, 3, \ldots, 10)$

λ	1	2	3	4	5	6	7	8	9	10
$e^{-\lambda}$	.36788	.13534	.04979	.01832	.006738	.002479	.000912	.000335	.000123	.000045

NOTA: Para obtener valores de $e^{-\lambda}$ para otros valores de λ, use las leyes de los exponentes

Ejemplo: $e^{-3.48} = (e^{-3.00})(e^{-0.48}) = (.04979)(.6188) = .03081.$

APÉNDICE H

Números aleatorios

51772	74640	42331	29044	46621	62898	93582	04186	19640	87056
24033	23491	83587	06568	21960	21387	76105	10863	97453	90581
45939	60173	52078	25424	11645	55870	56974	37428	93507	94271
30586	02133	75797	45406	31041	86707	12973	17169	88116	42187
03585	79353	81938	82322	96799	85659	36081	50884	14070	74950
64937	03355	95863	20790	65304	55189	00745	65253	11822	15804
15630	64759	51135	98527	62586	41889	25439	88036	24034	67283
09448	56301	57683	30277	94623	85418	68829	06652	41982	49159
21631	91157	77331	60710	52290	16835	48653	71590	16159	14676
91097	17480	29414	06829	87843	28195	27279	47152	35683	47280

ÍNDICE DE TEMAS

Aceptación, región de, 230
Ajuste de las distribuciones teóricas a distribuciones de frecuencias muestrales, 236, 257, 277
Aleatoriedad,
  completa, 345
  prueba de corridas para la, 375
Análisis combinatorio, 9, 20, 25, 32, 33
Análisis de varianza, 335-371
  a una vía, 335 (*véase también* Experimentos de un factor)
  con repeticiones o réplicas, 335, 345
  modelo lineal para el, 337
  no paramétrica, 374
  para experimentos de dos factores, 341, 345, 353
  para experimentos de tres factores, 362, 363
  para experimentos de un factor, 335, 337
  para números desiguales de observaciones, 339, 351
  tablas de, 339, 341
Aproximación normal a la distribución binomial, 120, 137, 159
Árbol, diagramas de, 9, 32
Axiomas de probabilidad, 6

Bernoulli, James, 117
Bloques
  aleatorios, 340, 345
  en experimentos de dos factores, 340, 345
Bondad de ajuste, 236
  en regresión, 286
  prueba chi cuadrado para la, 236-237

Cambio, 5
Cambio de variables, 47, 57, 74
Categorías, 173
Celdas, 238, 326, 327
Centro de gravedad, 287
Centroide *Véase* Centro de gravedad
Clases y frecuencias de clase, 173, 326, 327
Clasificación a tres vías, 362
Clasificación a una vía, 335 (*véase también* Experimentos de un factor)
Clasificación de experimentos a dos vías *Véase* Experimentos de dos factores
Coeficiente de correlación lineal 290, 310 (*véase también* Correlación)
  fórmula de producto–momento para, 291
  generalización del, 291, 313
  para datos agrupados, 327
  teoría de muestreo de, 291
Coeficiente de:
  contingencia, 239, 269, 280
  correlación 89, 100, 110 (*véase también* Correlación)
  correlación de rango *véase* Correlación de rango de Spearman
  correlación múltiple, 291, 313
  curtosis, 92
  determinación, 290
  sesgo, 92
Coeficientes binomiales, 11, 24-25, 33
Coeficientes de Fourier, 106
Combinaciones, 11, 23, 32
Complemento, 5
Confiabilidad, 209
Conjunto vacío, 4
Conjuntos medibles, 6
Contador Geiger, 159
Conteo, 9, 20, 32
  principio fundamental del, 9
Continua a la derecha, 41
Control de calidad, 235
Convoluciones, 48, 62, 63, 75
Corrección de Yates por la continuidad, 238, 260, 263, 266
Correlación, 285-334
  coeficiente de, 285, 290
  coeficiente generalizado de, 291, 313
  de la población, 293
  de rango, *Véase* Correlación de rango de Spearman
  fórmula de producto-momento para la, 291
  interpretación probabilística de la, 290, 316
  lineal, coeficiente de, 310
  lineal perfecta, 288, 290
  muestra, 290-291
  múltiple, coeficiente de, 291, 313
  para datos agrupados, 327
  prueba de hipótesis para la, 295
  tabla de, 327
  teoría muestral de la, 294, 320
  y dependencia, 295
Correlación de rango de Spearman, coeficiente de, 292, 314, 376, 389
Correlación y regresión lineal perfectas, 288, 290
Corrida, definición de, 375
Covarianza, 88, 100, 110
  teoremas sobre, 89
Cuadrado greco latino, 356, 358
Cuadrado latino, 346, 357
Cuadrados ortogonales, 346
Curtosis, 92, 105, 113
  de la distribución binomial, 118
  de la distribución de Poisson, 116
  de la distribución normal, 115
Curva cuadrática, 285
Curva de ajuste, 285 (*véase también* Correlación; Regresión)
Curva de potencia, 254, 270

Curva normal estándar, 119
áreas bajo la, 401
ordenadas de la, 400
Curva normal, 106 (*véase también* Distribución normal)
áreas bajo la, 133, 401
Curva parabólica, 285
Curvas de aproximación, 285
Curvas de mejor ajuste, 285
Curvas de operación característica, 235, 255, 277
Curvas de regresión no lineales, 291
Curvas OC *Véase* Curvas de operación característica

Datos originales, 173
Decil, 91
Decisiones estadísticas, 228
Desigualdad de Chebyshev, 90, 102, 111
Desviación estándar, 84, 92, 109
Desviación media, 92, 105
Determinación, coeficiente de, 290
Diagramas de árbol, *Véase* Árbol, diagramas de
Diagramas de dispersión, 285, 292, 301
Diseño de experimentos *Véase* Diseño experimental
Diseño experimental, 345
bloques aleatorios, 346
completamente aleatorio, 345
cuadrados greco latinos, 346, 358
cuadrados latinos, 346, 357
Dispersión *Véase* Medidas de dispersión
Distribución beta, 124, 145
Distribución bimodal, 91
Distribución binomial, 117, 127, 128, 157
aproximación de Poisson a la, 139
aproximación normal a la, 120, 137, 159
función generadora de momento de la, 118, 131, 132
media de la, 118, 132
negativa, 127
propiedades de la, 118
relación entre la distribución de Poisson y la, 120
relación entre la distribución normal y la, 120, 137
varianza de la, 118
Distribución chi cuadrado, 124, 146, 160, 212, 234
áreas bajo, 147-148
como un caso especial de la distribución gamma, 125
como un caso especial de la distribución normal, 147
función generadora de momento de la, 146
percentiles de la, 403
relacionada con la distribución normal, 147
teoremas sobre la, 125
Distribución condicional,49, 63
Distribución de Bernoulli, 117 (*véase también* Distribución binomial)
Distribución de Cauchy, 123, 144, 160
Distribución de frecuencia, 173, 188, 201
relativa, 173
Distribución de la muestra, 167
de desviaciones estándar, 172
de diferencias y sumas, 169-170, 182, 200
de las relaciones de varianzas, 171, 188, 201
de mediana, 172
de medias, 168, 172, 175, 199
de proporciones, 169, 172, 179, 199
de varianzas, 171, 172, 184, 200
Distribución de Maxwell, 128
Distribución de Pascal, 127
Distribución de Poisson, 120, 139, 159
ajuste de datos por la, 259
función generadora de momento de la, 121
media de la, 121
propiedades de la, 120
relación con la distribución binomial, 120
relación con la distribución normal, 121
Distribución de probabilidad continua, 41
Distribución de probabilidad empírica, 173
Distribución de Weibull, 128
Distribución exponencial, 128
Distribución *F*, 126, 150, 161, 212, 213
en análisis de varianza, 339
media, modo y varianza de la, 126
percentiles 95 para la, 404
percentiles 99 para la, 405
relación con las distribuciones chi cuadrado y *t*, 127
teoremas sobre la, 126-127
Distribución gamma, 124, 145, 160
fórmula asintótica de Stirling para la, 390
fórmula de recurrencia para la, 398
Distribución gaussiana *Véase* Distribución normal
Distribución geométrica, 127
Distribución hipergeométrica, 122, 143, 159
media y varianza de la, 122
relación con la distribución binomial, 122
Distribución multinomial, 122, 142, 159
Distribución normal, 118, 133, 158 (*véase también* Aproximación normal a la distribución binomial; Curva normal)
ajuste de datos por la, 252, 253
bivariada, 127, 153
función de densidad de la, 118
función generadora de momento de la, 120, 137
media de la, 119
ordenadas de la distribución normal estándar, 401
propiedades de la, 120
varianza de la, 120
Distribución normal bivariada, 127, 15
Distribución *t Véase* Distribución *t* de Student
Distribución *t* de Student, 125, 148, 160
áreas bajo la, 402
media de, 125
relación con la distribución normal, 126
relación con las distribuciones chi cuadrado y *F*, 127
valores percentiles de, 402
varianza de, 125
Distribución trimodal, 91
Distribución uniforme, 123, 143, 160
Distribuciones conjuntas, 44, 53
Distribuciones de probabilidad, 39, 41, 48, 50, 52, 72
de funciones de variables aleatorias, 47, 58, 73
discreta, 39
Distribuciones multimodales, 91
Ecuaciones no lineales reducibles a la forma lineal, 303
Ecuaciones normales, 287, 289, 323
para el plano de regresión de mínimos cuadrados, 289, 306
para la parábola de mínimos cuadrados, 289, 305
para la recta de mínimos cuadrados, 287, 296
Efectos
de bloques, 341
interacción, 343, 344

residual, 344
tratamiento, 340
Empate de registros *Véase* Empates
Empates, 373
prueba de Kruskal-Wallis corregida para, 374
Enfoque frecuentista de la probabilidad, 5
Error de tipo II, 229
Error de tipo I, 229
Error estándar del estimado, 289, 308
Error estándar, 167, 172
Error o residuo:
en análisis de varianza, 340, 343
en regresión, 286
Espacio muestral,3-4, 12
discreto, 4
finito, 4
infinito contable, 4
infinito no contable, 4
Espacios muestrales no discretos, 4
Esperanza, 82, 84, 93, 108, 292
condicional, 89, 101, 111, 292
de funciones de variable aleatorias, 83
definición de, 82
matemática, 82, 83
teoremas sobre, 84
Estadística de Fermi-Dirac, 34
Estadísticas de Bose-Einstein, 34
Estadísticas no paramétricas, 197, 374-376
Estadístico muestral, 167
Estimación, teoría de la, 209-227
de máxima verosimilitud, 213, 221, 225
eficiente, 209, 213, 223
en regresión, 286
por intervalo, 209
puntual, 209
Estimador
eficiente, 209, 213, 223
insesgado, 170, 209, 213, 223
Evento
cierto (seguro), 4
elemental (simple), 4
imposible, 4
simple, 4
Eventos
independientes, 9, 16, 31
mutuamente excluyentes, 5, 15
Expansión binomial, 25, 117
Expansión multinomial, 122
Experimento, aleatorio, 3, 12
Experimentos de tres factores, 362
Experimentos de dos factores, 340, 342
con replicación, 342, 354
tablas de análisis de varianza para, 343, 344
variaciones para, 340
Experimentos de Mendel, 262
Experimentos de un factor, 335, 346
métodos abreviados para, 336
notación para, 335
tabla para, 339
variaciones para, 336

Factorial, *n*, 10
aproximación de Stirling a, 12, 398
Fisher, R. A., 126, 213
Fórmula del producto–momento para el coeficiente de correlación lineal, 291, 310
Fórmulas de Euler, 99, 397
Fracaso, probabilidad de, 117
Frecuencia de celda, 326
de clase, 173, 188
esperada, 238
marginal, 238, 320
observada, 238
teórica, 238
Fuljan, 34
Función aleatoria (estocástica), *Véase* Variable aleatoria
Función beta, 124. 398
Función característica, 87-88, 99, 110
de la distribución binomial, 118
de la distribución de Poisson, 121
de la distribución normal, 120
teoremas sobre la, 88
Función complementaria de error, 399
Función de densidad, 41
condicional, 49
conjunta, 45, 73
de probabilidad, 41
de una suma de variables aleatorias, 48
marginal, 46
normal estándar, 119
transformada de Fourier, 88
Función de distribución, 39
conjunta, 45, 73
marginal, 46
para variables aleatorias continuas, 41, 52, 73
para variables aleatorias discretas, 40, 51, 72
Función de error, 119
complementaria, 399
Función de la distribución acumulada, 39
Función de operación característica, 253
Función de paso (escalonada), 41
Función de potencia,254
Función de probabilidad, 5, 39
condicional, 49
conjunta, 44
marginal, 46
Función escalonada, 41
Función factorial, 398
Función gamma, 124, 397
Función generadora de momento, 86-87, 97, 109
de la distribución binomial, 118, 132
de la distribución chi cuadrado, 146
de la distribución de Poisson, 121
de la distribución normal, 120, 137
teorema de la unicidad para, 87
teoremas sobre la, 87
Función monótonamente creciente, 41
Funciones de variables aleatorias, 83

Grados de libertad, 125, 126, 338, 342
Gráfico de polígono, 173
Gráficos de control de calidad, 235
Gráficos de control de calidad, 235, 257, 277
Gran media, 336, 340
Grupo de medias, 336

Hipótesis alterna, 228
Hipótesis estadísticas, 228
Hipótesis nula, 228

Inferencia estadística, 165
Integral de Fourier, 88
Integrales especiales, 398, 399
Intersección de conjuntos, 5
Intervalo de clase, 173, 188, 326, 327
Intervalos de confianza, 209-213
para diferencias y sumas, 211, 218, 224
para el cocientre entre varianza, 212, 220, 225
para medias, 210, 214, 216, 224
para proporciones, 211, 217, 224
para varianzas, 212, 219, 225
Invarianza bajo transformación, 290, 291
Inversión, fórmula de, 88

Jacobiano, 48, 57, 58, 60

Ley de los grandes números, 90-91. 103, 111, 118
para pruebas de Bernoulli, 118, 132
Ley fuerte de los grandes números, 91
Ley débil de los grandes números, 91
(*véase también* Ley de los grandes números)

Límite de clase, 173, 187, 188
Límites de confianza, 210 (*véase también* Intervalos de confianza)

Marca de clase, 173, 189
Media(s), 82, 103 (*véase también* Esperanza)
  aritmética, 82
  de fila, 336
  de grupo, 336
  de la distribución binomial, 118, 132
  de la distribución de Poisson, 121
  de la distribución *F*, 126
  de la distribución normal, 119
  de la distribución *t* de Student, 125
  de tratamientos, 336
  de un conjunto de números, 82
  de una muestra, 167, 190, 202
  distribución muestral de, 172
  general, 336, 340
  gran, 336
  muestral, 167, 190, 202
Mediana, 91, 103, 112
  distribución muestral de la, 172
  no unicidad de la, 91, 103, 104
Medidas de dispersión, 85, 92, 105
Medidas de tendencia central, 83, 91, 103, 112
Método de mínimos cuadrados, 286 (*véase* Mínimos cuadrados)
  para curvas de regresión, 287
  para parábolas, 287, 289, 306
  para rectas *Véase* Regresión líneal mínimos cuadrados
Métodos y fórmulas de notación, 174, 191, 193, 326, 327
Mínimos cuadrados:
  curva, 287
  curvas de regresión para, 287
  método de, 286
  parábola de, 287, 289, 305
  recta, 288, 295
  sentido de, 287
Moda, 91, 103, 112
  de la distribución beta 124
  de la distribución *F*, 126
Modelo matemático, 7
  para el análisis de varianza, 373, 341, 342
Momento(s), 86, 97, 109
  alrededor de la media, 86
  alrededor del origen, 86
  central, 86
  condicional, 89, 101, 111
  muestral, 190, 202
Moneda balanceada, 5
Muestras
  aleatorias, 166
  grandes, 210, 231
  independientes, 169, 182, 211
  pareadas, prueba del signo para, 373
Muestreo
  con remplazo, 122,
  sin remplazo, 122, 166

n factorial, 10
  aproximación de Stirling a, 12, 398
Nivel de significancia, 229, 230
Números aleatorios, 166
  tabla de, 407

Ocurrencias con probabilidades iguales, 5

Parábola, 285
  mínimos cuadrados, 287, 289, 305
Patrones
  cíclicos en la prueba de corridas, 375, 387
  de tendencia en pruebas de corridas, 375, 386
Percentiles, 91, 104, 112
Permutaciones, 10, 21, 31
Población, 165
  binomial, 166, 211
  con distribución normal, 166
  finita, 165, 168, 169
  infinita, 165, 166, 168
  multinomial, 236
  normal, 166
  parámetros de la, 166
  tamaño de la, 165
Poder de una prueba, 235, 252
Polígono de frecuencia, 173
Porcentaje de la distribución de frecuencia, 173
Potencia de una prueba *Véase* Poder de una prueba
Probabilidad, 5
  algunos teoremas importantes sobre, 6, 13
  asignación de, 7
  axiomas, 6
  condicional, 8, 31*teoremas sobre, 8
  de eventos, 117
  de fracaso, 117
  distribución de, 39
  empírica, 5, 173
  enfoque axiomático de la, 5
  enfoque clásico de la, 5
  enfoque frecuentista de la, 5
  función de la, 6, 39
  geométrica, 49, 65, 76
  papel gráfico de, 236, 257
  superficie de, 45
Probabilidades iguales, 7
Problema,
  de la aguja de Buffon, 70
  del cumpleaños, 30
  del sobre, 29, 30
Proporciones, distribución muestral de, 169, 172, 179, 199
Prueba(s),
  de Bernoulli, 117
  de cola derecha, 230
  de cola izquierda, 231
  de corridas para la aleatoriedad, 375, 386
  aplicaciones del, 376, 389
  prueba estadística de la media, 375
  prueba estadística de la varianza, 375
  de dos colas (bilaterales), 230
  de hipótesis, 228 *Véase* Pruebas de hipótesis y significancia
  de significancia, 229 (*véase también* Pruebas de hipótesis y significancia)
  de una cola, 230
  del signo, 372, 376
  independientes, 117
  no paramétricas, 372-396
  unilaterales, 230
Prueba
  chi cuadrado, 260, 279
  para la bondad de ajuste, 236-237, 260, 279
  F para medias iguales, 339
  H de Kruskal-Wallis, 374, 385
    corregida para empates, 374
  para aleatoriedad *Véase* Pruebas de corrida para la aleatoriedad
  U de Mann-Whitney, 373, 379
    prueba estadística de la media, 373
    prueba estadística de la varianza, 374
Pruebas de hipótesis y significancia, 228-284
  de la diferencia de medias, 232, 233, 245, 275
  de la diferencia de proporciones, 233, 245, 275
  de la diferencia entre coeficientes de correlación, 295
  del coeficiente de correlación, 295

para cocientes de varianzas, 234
para coeficientes de regresión en regresión lineal, 294
para medias, 232, 233, 239, 274
para muestras grandes, 231-233
para muestras pequeñas, 233-234
para proporciones, 232, 239, 274
para valores medios predichos en regresión lineal, 294
para valores predichos, 294
para varianzas, 234-235
que involucran la distribución chi cuadrado, 233, 250, 260, 276
que involucran la distribución *F*, 233, 251, 276
que involucran la distribución normal, 229
que involucran la distribución *t* de Student, 234, 248
Punto muestral, 3

Rango, semiintercuartílico, 92, 105, 112
Recta de mínimos cuadrados, 287-288
en términos de varianza y covarianza muestral, 288, 296, 318, 321
Región
crítica, 230
de aceptación, 230
de no significancia, 230
de rechazo, 230
Reglas de decisión, 229, 230
Regresión, 286
coeficiente de, 294, 319
curva de, 286, 291, 292, 293, 317
pendiente de la, 287
ecuación de, 286, 289
interpretación probabilística, 292, 316
múltiple, 289, 306
perfecta lineal, 288, 290
plano de, 289
superficies de, 289
teoría muestral de la, 294, 318
Regresión lineal, mínimos cuadrados, 289, 295, 318, 321
en términos de varianza y covarianza muestral, 288
Relación entre la teoría de la estimación y la prueba de hipótesis, 235
Relación entre las distribuciones binomial y de Poisson, 120
Relación entre las distribuciones binomial y normal, 120
Relación entre las distribuciones chi cuadrado, *t* y *F*, 127, 151
Relación entre las distribuciones normal y de Poisson, 121
Relación lineal, 286, 289, 313
Relación no lineal, 285, 313
Replicaciones, 335, 343-345

Sentido de mínimos cuadrados, 287
Serie de Fourier, 88, 106
Sesgo, 92, 105, 113

Significancia (*véase también* Prueba de hipótesis y significancia)
probable, 241-242, 249, 260
región de, 230
pruebas de, 228
Sumas especiales, 397

Tabla(s) de clasificación
a dos vías, 237, 343, 344
de frecuencia, 173 (*véase también* Distribución de frecuencia)
de probabilidad conjunta, 44
doble, 238
simple, 237, 339
Tabla(s) de:
análisis de varianza
a dos vías, 341
a una vía, 339
áreas bajo la curva normal estándar, 401
contingencia, 237, 265, 279
números aleatorios, 407
ordenadas de la curva normal estándar, 400
percentiles 95 para la distribución *F*, 404
percentiles 99 para la distribución *F*, 405
percentiles para la distribución chi cuadrado, 403
percentiles para la distribución *t* de Student, 402
valores de $e^{-\lambda}$, 406
Tamaño muestral, 165
Tendencia central *Véase* Medidas de tendencia central
Teorema de Bayes (regla), 9, 19-20, 32
Teorema del límite central, 121, 141, 168, 272
para variables aleatorias con distribución binomial, 141
demostración de, 142
Teoría muestral:
de correlación, 294, 320
de la regresión, 294, 318
Total de frecuencia marginal, 320
Transformación Z de Fisher, 295
en análisis de varianza, 335
en correlación, 295
Transformada de Fourier, 88
Transformada inversa de Fourier, 88
Traslación de ejes, 287
Tratamientos, 35, 340

Unidades estándar, 86, 93
Unión de conjuntos, 4

Valor esperado, 82 (*véase también* Esperanza)
Valor estándar, 86, 119
Valor *P*, 226
Valores críticos, 210
Valores exponenciales, tabla de, 406
Variable aleatoria (estocástica), 38
absolutamente continua, 41
asintóticamente normal, 120, 121 168
con distribución beta, 124
con distribución binomial, 117
con distribución chi cuadrado, 124
relación con las distribuciones y *F*, 127, 151
con distribución de Bernoulli, 117
con distribución de Cauchy, 123
con distribución de Poisson, 120
con distribución gamma, 124
con distribución normal, 119
continua, 41, 52, 72
dependiente, 46
discreta, 38, 50, 72
estandarizada, 86
funciones de, 83
independiente, 46, 47, 53, 73
no correlacionada, 89
no discreta, 38, 50
uniforme, 123
Variable dependiente, 286, 289, 297
Variable estocástica *Véase* Variable aleatoria
Variable independiente, 286, 289, 297, 298
Variables transformadas, 285
Variación (*véanse también* Variación explicada; Variación total; Variación no explicada)
aleatoria, *Véase* Variación residual o al azar
debida al error, 341
distribuciones de la, 338
entre bloques, 344
entre tratamientos, 336, 344

entre tratamientos, 337
explicada, 290, 293, 313, 314
grados de libertad dentro de tratamientos, 339
grados de libertad para el residuo, 344
grados de libertad para el total, 339, 344
grados de libertad para entre bloques, 344
grados de libertad para entre tratamientos, 339, 344
métodos abreviados para la, 336
no explicada, 290, 293, 313, 314
para experimentos de dos factores, 340
residual o al azar, 341, 344
en análisis de varianza, 343, 344
en regresión, 285,
total, 290, 293, 310, 311, 313, 314, 340, 344
valores esperados para la, 337
Varianza, 84-85, 95, 109
combinada, 234
como segundo momento, 86
condicional, 89, 101, 111
de la distribución binomial, 118
de la distribución F, 126
de la distribución muestral de medias, 168
de la distribución normal, 120
de la distribución *t* de Student, 125
de la población, caso en que se desconoce la, 171, 187
muestral, 170, 190, 202
teoremas sobre la, 85

ÍNDICE DE PROBLEMAS RESUELTOS

Ajuste de datos por distribuciones teóricas, 257
Aleatoriedad, prueba de corridas para, 386
Análisis combinatorio, 20
  usando probabilidad, 25
Aproximación normal a la distribución binomial, 137

Bondad de ajuste, 259, 264

Cálculo de probabilidades, 14
Cambio de variables, 57, 69
Clasificación
  a dos vías, 352
  simple, 346
  tridimensional, 362
Coeficiente de:
  contingencia, 269
  correlación lineal, 310
  correlación, 100
  determinación, 322
  (*Véanse también* Contingencia; Correlación)
Coeficiente de correlación
  generalizado, 313
  lineal (*véase* Correlación)
  múltiple, 314
  poblacional, 113
Coeficiente de determinación, 322
Coeficientes binomiales, 24-25
Coeficientes de Fourier, 106
Combinaciones, 23
Conteo, 20
Contingencia, coeficiente de, 269
Convoluciones, 62
Correlación:
  coeficiente de, 100, 153
  de rango, 314, 390
  fórmula de producto–momento para la, 312
  generalizada, 313
  interpretación probabilística de la, 316
  lineal, 310
  múltiple, 314
  tabla, 327
  teoría muestral de la, 320
Covarianza, 100,
Cuadrados greco latinos, 358
Cuadrados latinos, 357
Curtosis, 105, 194
Curvas de operación característica, 252, 270

Densidad condicional, 64, 66
Desigualdad de Chebyshev, 102
Desviación estándar, 95
Diagrama de dispersión, 301
Diagramas de árbol, 20
Dispersión, medidas de, 105
Distribución:
  beta, 145
  binomial, 128
    aproximación de Poisson a la, 139
    aproximación normal a la, 137, 141
    función generadora de momento de la, 132
  chi cuadrado:
    función generadora de momento de la, 146
    pruebas que involucran la, 250, 260
    relación con la distribución normal, 147
    relación con las distribuciones *F* y *t*, 151
  condicional, 63
  conjunta, 53
  continua, 52
  de Cauchy, 144-145
    función característica de, 144
    función generadora de momento de la, 144
    relación con la distribución uniforme, 145
  de frecuencia, 188
  de las relaciones entre *F*, chi cuadrado y *t*, 151
  de las relaciones entre varianzas, 188
  de medias, 175
  de Poisson, 139
    función generadora de momento de la, 140
  de proporciones, 179
  de Weibull, 153
  discreta, 4
  F, de Fisher, 150
    relación con las distribuciones chi cuadrado y *t*, 151
    pruebas que involucran la, 251
  gamma, 145
  hipergeométrica, 143
  marginal, 53
  muestral, 175, 182, 184, 188
    de diferencias y sumas, 182
  multinomial, 142
  normal, 133
    aproximación a la distribución binomial, 137
    bivariada, 153
  t de Student, 148
    relación con las distribuciones *F* y chi cuadrado, 151
    pruebas que involucran la, 248
  uniforme, 143
Distribuciones conjuntas, 53
Distribuciones de probabilidad:
  continua, 52
  discreta, 50
Distribuciones muestrales:
  de diferencia de medias, 182
  de la media cuando se desconoce la varianza de la población, 187
  de las relaciones entre varianzas, 188
  de medias, 175, 196
  de proporciones, 179
  de suma de medias, 182
  de varianzas, 184
Distribuciones teóricas, ajuste de datos por, 257

Ecuaciones no lineales reducibles a la forma lineal, 303
Error estándar del estimado, 308
Espacios muestrales, 12
Esperanza de variables aleatorias, 93
  condicional, 101

Estimación del intervalo de confianza:
de la media cuando se desconoce la varianza de la población, 187
para diferencias en medias, 218
para diferencias en proporciones, 218
para el cociente entre varianzas, 220
para la desviación estándar, 223
para medias en muestras grandes, 214
para medias en muestras pequeñas, 216
para proporciones, 217, 222
para varianzas, 219
Estimadores:
de máxima verosimilitud, 221
para la media de la distribución normal, 221
para la varianza de la distribución normal, 222
eficientes, 213
insesgados, 213
intervalo de confianza *Véase* Estimación del intervalo de confianza
Eventos, 12
independientes, 16
mutuamente excluyentes, 15
Experimentos aleatorios, 12
Experimentos de dos factores, 35
con replicación, 354
Experimentos de tres factores, 362
Experimentos de un factor, 346

Fórmula del producto–momento, 312
Frecuencia:
distribuciones de, 188
histogramas de, 189
polígonos de, 188
Función característica, 99, 106
de la distribución de Cauchy, 144
Función de densidad
condicional, 63, 66
conjunta, 54
marginal, 64
Función de distribución
continua, 52
discreta, 51
marginal, 53, 67
Función de operación característica, 254
Función de potencia, 258
Función generadora de momento, 97, 109
de la distribución binomial, 132
de la distribución de Cauchy, 144
de la distribución de Poisson, 140
de las sumas de variables aleatorias independientes, 97

Gráficos de control de calidad, 256

Interpretación probabilística de la regresión y la correlación, 316

Ley de los números grandes, 103
para pruebas de Bernoulli, 132

Media muestral:
cálculo de la, 190
distribución de la, 175
fórmula de notación para la, 191
Mediana, 104
Medidas de dispersión, 105
Medidas de la tendencia central, 103
Mínimos cuadrados:
parábola de, 305
recta de, 295
Moda, 104
Modificaciones para números desiguales de observaciones, 351
Momento, 97
cálculo de, 101
condicional, 101
para muestras, cálculo de, 190

Papel log-log, 304
Patrón cíclico en pruebas de corridas, 387
Patrón de dispersión, en pruebas de corridas, 387
Percentiles, 105
Permutaciones, 10, 21
Probabilidad geométrica, aplicaciones con, 65
Probabilidad:
asignación de, 7
cálculo de la, 14
usando el análisis combinatorio, 25
condicional, 16
distribuciones de, 50
geométrica, 65
teoremas sobre, 6, 13
Problema de la aguja de Buffon, 70
Proporciones, distribución muestral de, 179
Prueba(s):
de diferencia de medias, 245
de diferencia de proporciones, 245
para medias usando distribuciones normales, 239
para proporciones usando distribuciones normales, 239
que involucran la distribución chi cuadrado, 250, 260
que involucran la distribución *F*, 251
que involucran la distribución *t* de Student, 248
Prueba chi cuadrado para la bondad de ajuste, 260, 264, 271
Prueba de corridas para aleatoriedad, 386
aplicaciones de, 389
Prueba de hipótesis *Véase* Pruebas
Prueba del signo, 376
Prueba *H* de Kruskal-Wallis, 385
Prueba *U* de Mann-Whitney, 379
distribución muestral de la, 383

Regresión múltiple, 306
Regresión:
interpretación probabilística de la, 316
múltiple, 306
recta de mínimos cuadrados para la, 295
teoría muestral de la, 318
Relación lineal, 313
Relación no lineal, 313

Series de Fourier, 106
Sesgo, 105, 194

Tabla de correlación, 327
Tablas de contingencia, 265
Tendencia central, medidas de, 103
Teorema de Bayes, 19-20
Teorema del límite central, 141
demostración del ,142
para variables aleatorias binomiales, 141
Teoría muestral de:
correlación, 320
de varianzas, 184
de Weibull, 153
regresión, 318
uniforme, 143

Valores de tendencia, 323
Variables aleatorias
continuas, 52
discretas, 50
esperanza condicional de, 101
esperanza de, 93
independientes, 52
Variables, cambio de, 57, 70
Variación:
explicada, 311
no explicada, 314
total, 311
Varianza, 95
condicional, 101
de la muestra:
fórmula de notación para la, 192
cálculo de la, 190
distribución muestral de la, 184

Schaum

Cada año, cientos de miles de estudiantes prefieren la Serie Schaum porque produce resultados y lo comprueban al mejorar sus promedios con esta indispensable guía.

Si no posees tiempo suficiente, pero buscas la excelencia en clase, esta obra te ayudará a:

- Prepararte para tus exámenes
- Encontrar respuestas con rapidez
- Estudiar mejor y más efectivamente
- Captar la idea sin consultar libros más extensos

La serie Schaum te da la información que tus profesores esperan que tengas a la mano, pero sin detalles innecesarios. Además tiene suficientes ejercicios para que pruebes tus habilidades.

Compatible con cualquier libro de texto, te deja estudiar a tu propio ritmo, resaltando todos los puntos importantes que necesitas recordar.

En este libro encontrarás:

- 760 problemas totalmente desarrollados con soluciones paso a paso
- Cientos de problemas adicionales de práctica con sus respuestas
- Cobertura de todas las bases del curso
- Metodología fácil de entender

Si deseas obtener las mejores calificaciones y una comprensión completa de la probabilidad y la estadística, ¡esta poderosa herramienta de estudio es el mejor tutor que puedes tener!

Incluye capítulos sobre:
Probabilidad básica / Variables aleatorias y distribuciones de probabilidad / esperanza matemática / distribuciones de probabilidad especial / teoría de muestreo / teoría de estimación / pruebas de hipótesis y significación / ajuste a la curva, regresión y correlación / análisis de varianza / pruebas no paramétricas.

Mc Graw Hill **McGraw-Hill Interamericana**

The McGraw-Hill Companies

ISBN 970-10-4231-X

9 789701 042311

Visite nuestra página WEB
www.mcgraw-hill.com.mx